目　录

第1章　生物信息学概论 …… (1)
1.1　生物信息学的定义 …… (1)
1.2　生物信息学的发展历史 …… (2)
1.3　生物信息学的主要研究领域、基本问题和方法 … (4)
1.3.1　生物学数据库的建立和搜寻 …… (4)
1.3.2　DNA 和蛋白质的序列分析 …… (5)
1.3.3　基因结构的预测 …… (6)
1.3.4　蛋白质结构和功能的预测 …… (6)
1.3.5　基因组数据的分析 …… (7)
1.3.6　比较基因组和系统发生遗传学分析 …… (7)
1.3.7　功能基因组和蛋白组学数据的分析 …… (8)
1.3.8　信号传导、代谢和基因调节途径的构建与描述 …… (8)
1.4　生物信息学今后的发展方向和趋势 …… (8)
1.5　生物信息学家应具备的基本知识和技能 …… (9)
第2章　生物信息的流向 …… (13)
2.1　生物信息在细胞内的流向 …… (13)
2.1.1　基因组:DNA 和基因的结构 …… (14)
2.1.2　转录组:mRNA 的结构 …… (17)
2.1.3　蛋白组:蛋白质的结构、相互作用和定位 …… (19)
2.1.4　代谢组:生化代谢途径和信号传导途径 …… (24)
2.1.5　细胞组:细胞器的结构与功能 …… (28)
2.2　生物信息在细胞间的流向 …… (33)
2.2.1　细胞周期及其调控 …… (33)
2.2.2　有丝分裂 …… (34)
2.2.3　减数分裂 …… (34)
2.3　生物信息在世代间的传递 …… (35)
2.3.1　孟德尔遗传方式 …… (35)
2.3.2　多基因遗传 …… (39)
2.4　生物信息流向紊乱与遗传病 …… (41)

2.4.1 基因突变的定义和分类 …… (41)
2.4.2 基因突变的效应 …… (43)
2.4.3 遗传病的基本概念和分类 …… (44)
第3章 人类基因组与基因组生物信息学 …… (47)
3.1 人类基因组计划:基因组的解剖结构 …… (49)
3.1.1 人类基因组计划的实施过程 …… (49)
3.1.2 后基因组时代的任务和目标 …… (52)
3.1.3 人类基因组的结构特点 …… (53)
3.1.4 人类基因组的多样性 …… (54)
3.2 人类蛋白组计划 …… (55)
3.2.1 血浆蛋白组计划 …… (55)
3.2.2 肝脏蛋白组计划 …… (55)
3.2.3 蛋白组标准化数据库 …… (55)
3.3 比较基因组学 …… (56)
3.4 基因组医学及相关的社会、伦理、法律问题 …… (56)
3.4.1 预测医学的发展趋势 …… (56)
3.4.2 应用基因组学的发展 …… (57)
3.4.3 遗传咨询与遗传伦理学的新范畴 …… (59)
3.5 基因组图、全基因组扫描和定位克隆 …… (59)
3.5.1 基因组图 …… (59)
3.5.2 全基因组扫描 …… (60)
3.5.3 定位克隆 …… (62)
3.6 基因组的组装与分析 …… (64)
3.6.1 基因组图谱的整合 …… (64)
3.6.2 基因在基因组中的定位方法 …… (66)
3.6.3 基因的结构组成与序列检索 …… (67)
3.6.4 定位克隆中关键区的定位和候选基因的筛选 …… (70)
3.6.5 同系基因和同源基因检索 …… (72)
第4章 生物信息学常用数据库介绍 …… (75)
4.1 分子生物学数据库 …… (75)
4.1.1 分子生物学数据库发展简史 …… (75)
4.1.2 分子生物学数据库分类 …… (76)
4.2 序列数据库 …… (78)
4.2.1 核苷酸序列数据库 …… (78)
4.2.2 蛋白质序列数据库 …… (81)
4.3 基因组图谱数据库 …… (84)
4.3.1 基因组图的组成成分和类型 …… (85)

4.3.2 定位数据库 …… (85)
4.4 医学数据库 …… (91)
4.4.1 OMIM …… (92)
4.4.2 HGMD …… (92)
4.4.3 其他突变数据库 …… (93)
4.5 生物信息的获取与数据库查找工具 …… (94)
4.5.1 Entrez …… (95)
4.5.2 FASTA …… (96)
4.5.3 BLAST …… (96)
附录4-1 EMBL Flatfile 格式 …… (100)
附录4-2 GenBank Flatfile 格式 …… (102)
附录4-3 Swiss-Prot Flatfile 格式 …… (104)
第5章 DNA 和蛋白质的序列生物信息学 …… (111)
5.1 成对序列比对 …… (112)
5.1.1 编辑距离 …… (112)
5.1.2 点阵描图 …… (113)
5.1.3 记分模型 …… (114)
5.1.4 替换记分 …… (116)
5.1.5 最佳比对 …… (118)
5.2 多序列比对 …… (123)
5.2.1 多序列比对主要方法和常用软件工具 …… (123)
5.2.2 渐进整体比对法 …… (124)
5.2.3 反复整体比对法 …… (126)
5.2.4 遗传算法 …… (127)
5.2.5 局部比对 …… (127)
5.2.6 序列比对中的统计方法 …… (128)
5.2.7 MSA 的编辑和修改 …… (131)
5.3 多序列比对在生物信息学研究中的应用 …… (131)
5.3.1 序列结构域和基序的寻找 …… (131)
5.3.2 基因调节因子预测 …… (132)
5.3.3 基因组组装 …… (132)
5.3.4 系统发生遗传学分析 …… (133)
第6章 蛋白质结构分析与结构生物信息学 …… (137)
6.1 蛋白质基本特性分析 …… (137)
6.2 蛋白质结构域、基序与结合部位分析 …… (138)
6.3 蛋白质拓扑结构、折叠和三维结构模型 …… (138)
6.3.1 同源性模型 …… (139)

6.3.2 折叠识别 …… (140)
6.3.3 *ab initio* 折叠 …… (140)
6.4 蛋白质结构视觉和分子模拟 …… (141)
6.4.1 蛋白质结构视觉 …… (141)
6.4.2 蛋白质分子模拟 …… (142)
6.5 蛋白质相互作用分析 …… (142)
6.5.1 进化遗传分析 …… (143)
6.5.2 *ab initio* 预测 …… (144)
6.5.3 贝叶斯网络法 …… (144)
6.6 蛋白质结构分析的基本流程 …… (145)
6.7 蛋白质结构分析的技术平台 …… (147)
6.7.1 X－衍射蛋白质晶体结构分析 …… (147)
6.7.2 核磁共振光谱分析 …… (147)
6.7.3 冷冻电镜技术 …… (148)
6.8 常用蛋白质结构数据库介绍 …… (148)
6.8.1 PDB …… (148)
6.8.2 SCOP 和 CATH …… (149)
6.8.3 Pfam …… (151)
第7章 微阵列生物信息学 …… (153)
7.1 微阵列简介 …… (153)
7.2 微阵列的主要类型 …… (154)
7.2.1 cDNA 微阵列 …… (154)
7.2.2 高密度寡核苷酸阵列 …… (156)
7.3 微阵列的应用 …… (158)
7.4 微阵列实验设计 …… (159)
7.4.1 对照组的选择 …… (159)
7.4.2 重复样本的使用 …… (160)
7.4.3 随机化原则 …… (161)
7.5 数据处理 …… (162)
7.5.1 丢失数据和极端值的处理 …… (162)
7.5.2 数据的正态性和线性检查 …… (163)
7.5.3 数据标准化 …… (164)
7.6 数据分析 …… (167)
7.6.1 差异表达分析 …… (167)
7.6.2 主成分分析 …… (170)
7.6.3 聚类分析 …… (170)
第8章 进化遗传学和统计遗传学 …… (175)
8.1 分子进化与遗传距离 …… (175)

8.1.1 进化钟和蛋白质演变速率 …… (175)
8.1.2 核酸分子的演变和模型 …… (176)
8.1.3 遗传距离的计算方法 …… (177)
8.2 进化遗传模型 …… (178)
8.3 遗传树的构建 …… (178)
8.3.1 建立数据模型 …… (179)
8.3.2 决定取代模型 …… (179)
8.3.3 建树方法 …… (180)
8.3.4 遗传树搜索 …… (182)
8.3.5 确定树根 …… (183)
8.3.6 评估进化树 …… (183)
8.4 MCMC 方法在进化遗传分析中的应用 …… (183)
8.4.1 贝叶斯定理 …… (183)
8.4.2 MCMC 的基本概念 …… (184)
8.4.3 MCMC 的建树方法 …… (184)
8.4.4 MCMC 的难点 …… (185)
8.5 疾病基因的连锁与关联分析 …… (185)
8.5.1 家族聚集性分析 …… (187)
8.5.2 LOD 分析 …… (188)
8.5.3 TDT 分析与 ASP 分析 …… (189)
8.5.4 病例对照资料分析 …… (190)
8.5.5 QTL 定位 …… (191)
8.6 群体遗传学分析方法 …… (192)
8.6.1 Hardy – Weinberg 平衡 …… (192)
8.6.2 近婚系数 …… (193)
8.6.3 选择作用 …… (195)
8.6.4 基因突变 …… (196)
8.6.5 遗传漂变 …… (196)
8.7 遗传风险计算 …… (197)
8.7.1 单基因病再发风险率的计算 …… (197)
8.7.2 多基因病再发风险率的计算 …… (199)
8.7.3 染色体病再发风险率的计算 …… (199)
8.8 亲子关系排除与肯定概率 …… (200)
第9章 药物基因组学与化学信息学 …… (203)
9.1 药物个体反应性的遗传基础 …… (204)
9.1.1 药物遗传学疾病概述 …… (205)
9.1.2 药物代谢酶类 …… (207)
9.1.3 转运蛋白 …… (207)

9.1.4 药物受体 …… (208)
9.2 药物遗传性状的鉴定方法 …… (208)
9.2.1 候选基因研究中的病例对照分析 …… (209)
9.2.2 全基因组连锁不平衡作图 …… (210)
9.2.3 药物基因组学研究方法的设计 …… (211)
9.2.4 药物基因组学常用技术平台 …… (212)
9.2.5 个体化医学在诊断和治疗中的应用 …… (213)
9.3 药物筛选与计算机辅助药物设计 …… (215)
9.3.1 药物研发的基本过程 …… (215)
9.3.2 药物锚定方法 …… (217)
9.3.3 计算机辅助药物设计 …… (218)
第10章 统计生物信息学 …… (221)
10.1 概率论基础 …… (221)
10.1.1 随机变量 …… (222)
10.1.2 概率分布函数和频数分布 …… (222)
10.2 统计学基本概念 …… (226)
10.2.1 均数与方差 …… (226)
10.2.2 置信区间 …… (229)
10.2.3 假设检验 …… (231)
10.2.4 Ⅰ类和Ⅱ类错误 …… (232)
10.3 常用统计检验方法 …… (233)
10.3.1 t 检验 …… (234)
10.3.2 方差分析 …… (237)
10.4 关联分析 …… (239)
10.4.1 相对危险度 …… (239)
10.4.2 几率比值 …… (240)
10.4.3 归因危险度 …… (241)
10.4.4 率的标准化 …… (243)
10.4.5 几率比值的校正 …… (244)
10.5 相关与回归分析 …… (246)
10.5.1 线性相关与回归 …… (246)
10.5.2 协方差分析 …… (247)
10.5.3 多元线性回归 …… (248)
10.5.4 Logistic 回归 …… (250)
10.6 常用统计软件介绍 …… (253)
10.6.1 SPSS 软件 …… (253)
10.6.2 SAS 软件 …… (254)

第1章 生物信息学概论

1.1 生物信息学的定义

近年来在基因组学(genomics)和生物信息学(bioinformatics)领域所发生的极为迅速的革命性进展,正在对生物医学产生空前的影响。生物信息学这一术语也因此走出科学论文和学术会议场合而成为从新闻媒体,到公司董事会议室,乃至到政治高峰会谈中经常提到的时髦科学词语。作为一门新兴的学科,生物信息学和与其紧密相关的基因组学也无疑迅速成为最热门的学科。本书的目的是为有兴趣了解和有志从事生物信息学研究的读者提供一个对此学科较为系统的介绍。

生物信息学作为一门新兴的学科,其广义的定义泛指任何与运用计算机及其网络和数据库进行大规模生物学数据的收集、组织管理和分析相关的研究领域,而较为狭义的定义指用计算机的手段和方法来对生物遗传的信息和数据进行管理和分析。许多时候,生物信息学与计算生物学(computational biology)被等同使用。但精确地说来,计算生物学指的是将计算机系统和计算机运算法则应用于生物现象的模型研究。生物信息学以计算机技术为研究手段和工具,同时采用数学、统计学的模型、模拟研究手法来解决生物科学的问题,因而成为生物学、统计学、数学、计算机学甚至工程学等多学科之间的交叉领域(图1-1)。其发展一方面依赖于这些学科的发展,但同时又对这些学科提出新的要求和挑战,为其开辟新的研究领域而促进其发展。

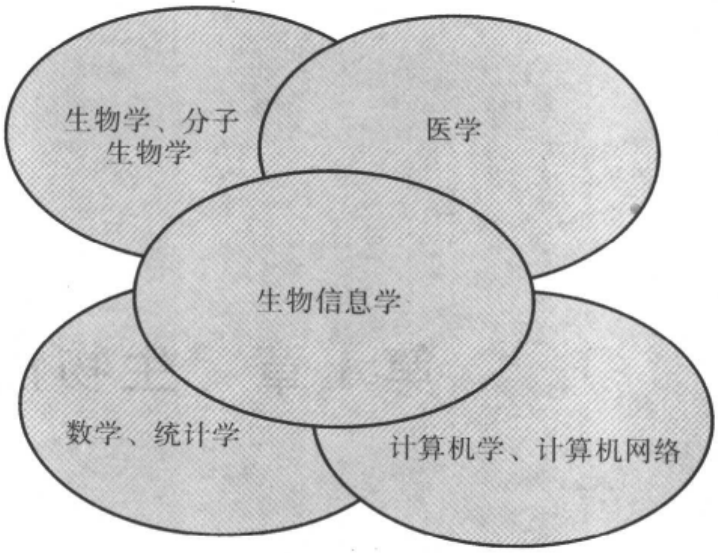


图 1-1　生物信息学与其他学科之间的关系

1.2　生物信息学的发展历史

生物信息学早期的研究对象主要限于 DNA 序列的储存和分析，而其最近的迅速发展主要缘于基因组计划中大规模 DNA 数据管理和分析的需要。其作为一门独立的学科只有近 20 年的历史，而这一名词的准确起源似乎无可考证。但事实上，与生物信息学相关的研究可以追溯到远至上世纪中期对蛋白质和 DNA 结构预测的模型研究。为帮助读者更好地了解生物信息学的起源和发展历史，以下简要列出与生物信息学发展相关的在生物学、计算机学和网络技术领域的重要里程碑。

1933 年：Tiselius 发明用于分离可溶蛋白质的电泳技术。

1951 年：Pauling 和 Corey 提出蛋白质的 α - 螺旋和 β - 片层结构。

1953 年：Waston 和 Crick 根据 Franklin 和 Wilkins 的 X - 晶体衍射数据提出 DNA 双螺旋结构[1]。

1954 年：Sanger 发表牛胰岛素的蛋白序列。

1958 年：Texus Instrument 公司发明第一个整合电路，美国 The Advanced Research Project Agency（ARPA）成立。

1969 年：由位于斯坦福大学、犹他大学、加州大学洛杉矶和圣巴分校的计算机相连的 ARPA 网络（ARPANET）建成。

1970 年：用于序列比较的 Needleman - Wunsch 运算法则发表。

1971 年：Ray Tomlinson 发明电子信函程序。

1972 年：Paul Berg 研究组创立第一个重组 DNA 分子。

1973 年：Brookhaven 蛋白数据库（Protein Data Bank，PDB）宣布成立。同时 Robert Metcalfe 以描述网络卡（Ethernet）的论文获哈佛大学博士学位。

1974 年:Vint Cerf 和 Robert Kahn 建立以计算机相连的网络构成的英特网(Internet)概念和发展传输控制方法(transmission control protocol, TCP); Charles Goldfarb 发明 SGML (standardized general markup language)格式。

1975 年:Bill Gates 和 Paul Allen 创立微软公司(Microsoft Corporation);O'Farrell 发表 SDS-聚丙烯酰胺和等电聚集相结合的二维电泳分离蛋白质技术;Southern 发表 DNA 与 DNA 的杂交技术(Southern blot)。

1976 年:贝尔实验室发明 Unix-To-Unix 复制方法(UUCP)。

1977 年:发表有关 Brookhaven 蛋白数据库的全面描述[2];哈佛大学的 Allan Maxam 和 Walter Gilbert 及英国医学研究委员会的 Frederick Sanger 分别发表 DNA 测序方法。

1978 年:Tom Truscott, Jim Ellis 和 Steve Bellovin 建立第一个连接 Duke 和北卡大学的 Usenet。

1980 年:第一个完整的生物基因序列(FX174)发表,此基因包含 5 386 个核苷酸碱基对,编码 9 个蛋白质;Wuthrich 等发表多维 NMR 蛋白结构测定技术的详细描述[3];IntelliGenetics 成立并发展用于蛋白质和 DNA 序列分析的 IntelliGenetics 软件包。

1981 年:用于序列比对的 Smith-Waterman 运算法则发表;IBM 公司向市场推出个人计算机。

1982 年:Wsiconsin 大学生物工程中心属下的遗传计算组 Genetics Computer Group (GCG)成立,其主要产品是有名的 GCG 分子生物学计算工具。

1984 年:Macintosh 公司的苹果计算机问世;Jon Postel 的 Domain Name System (DNS)开始在网络上使用。

1985 年:FAST$_p$ 序列分析运算法则发表;Kary Mullis 和其同事发明多聚酶链反应 (polymerase chain reaction, PCR)。

1986 年:Thomas Roderick 首次使用基因组学(genomics)作为一种新的杂志名用来描述有关基因组物理构图、序列测定和分析基因的学科;日内瓦大学和欧洲分子生物学实验室(EMBL)创立 SWISS-PROT 蛋白序列数据库。

1987 年:酵母人工染色体(yeast artificial chromosome, YAC)的使用被发表[4];大肠杆菌的物理图谱被发表[5];麻省理工学院的 Larry Wall 创立 PERL 语言(practical extract report language)。

1988 年:美国国立卫生研究所下属的国家生物工程信息中心(National Center for Biotechnology Information, NCBI)和欧洲生物信息学研究所(European Bioinformatics Institute, EBI)成立;Pearson 和 Lupman 发表 FASTA 运算法则。

1990 年:人类基因组计划开始,计划用 30 亿美元的预算在 15 年的时间内完成人类 30 亿碱基对的测序和基因确定;BLAST 程序建立; Tim Berners-Lee 发表第一份 HTML 文件。

1991 年:日内瓦研究所 CERN 宣布建立用于万维网(World Wide Web, WWW)的规程;Linus Torvalds 宣布后来成为 LINUX 的类 UNIX 操作系统;Craig Venter 在 Science 杂志上描述表达序列标签(expressed sequence tag, EST)的建立和使用[6]。

1992 年:Craig Venter 在美国马里兰州成立基因组研究所 (The Institute of Genome Research, TIGR),成为细菌基因组测序研究的先驱;Mel Simon 和同事宣布细菌人工染色体

(bacterial artificial chromosome, BAC)在 DNA 克隆中的应用。

1994 年:网景公司(Netscape)成立,发放网络浏览器 Navigator。

1995 年:微软公司发布第一版 Internet Explorer (IE);太阳(SUN)和网景公司(Netscape)联合发布第一代 Javascript;网站 SERVER 程序 APACH 被发布;TIGR 使用其创立的全基因组鸟枪法(whole genome shotgun, WGS)完成流感嗜血杆菌和解脲支原体两种细菌基因组,标志着基因组时代的开始。

1996 年:W3C 公司发布初级 XML 版本;第一个真核生物基因组——面包酵母的测序开始; Bairoch 等报道 Prosite 数据库;Affymetrix 推出第一代基因芯片(DNA Chips)。

1997 年:第一个实验模式生物 - 大肠杆菌的基因组被完成。

1998 年:面包酵母和线虫基因组发表。

2000 年:第一个植物拟南芥基因组和果蝇基因组完成。

2001 年:国际人类基因协作组和 Craig Venter 领导的 Celera 公司分别在《自然》和《科学》杂志同时发表人类基因组草图[7,8]。

2002 年:小鼠、水稻基因组工作草本序列发表。

2003 年:人类基因组计划完成。

从以上生物信息学发展史可以看出,其发展与计算机及网络技术的发展紧密相连。其中最为关键的包括计算机速度、内存容量、数据储存容量的快速提高,网络的普及和网络工具的发展。事实上在过去的 10 年中,计算机的计算速度和生物数据的信息量差不多以指数速度同步增长。

1.3 生物信息学的主要研究领域、基本问题和方法

目前的生物信息学研究,已从早期以数据库的建立和 DNA 序列分析为主的阶段,转移到后基因组学时代以比较基因组学(comparative genomics)、功能基因组学(functional genomics)和整合基因组学(integrated genomics)为中心的新阶段。生物信息学的研究领域也迅速扩大。以下简要归纳当前生物信息学研究中的基本问题。本书的其他章节将对其中大多数的问题分别进行详细讨论。

1.3.1 生物学数据库的建立和搜寻

生物学数据库贮存生物信息学研究的原始数据,是生物信息学存在和发展的基础。从 Dayhoff 及其同事 20 世纪 60 年代建立第一个已知蛋白质序列的分子生物学数据库到今天经历了突飞猛进的发展。80 年代 GenBank、EMBL(European Molecular Biology Laboratory)和 DDBJ(DNA DataBank of Japan) 以 DNA 序列为主的世界三大标准数据库的建立为分子生物学数据库的发展奠定了基础,并发挥了核心作用。计算机网络的发展与迅速普及和使用极大地促进了数据库的发展,并保证其数据为广大的用户方便地获取,而计算机储存媒体的发展和储存量的快速增长满足了生物数据指数增长对储存媒体的需求。其中同样关键的是关系数据库技术(relational database)的发展促进了对数据库的使用。多年前在所有的分子生

物学学术相关杂志中确立统一标准，要求所有新发表的分子序列在正式发表之前必须储存到 GenBank、EMBL 或 DDBJ 中的任何一个数据库并获得一个统一的序列登记号码（accession number）。这对分子生物学序列数据的标准化和保证数据库所含数据的公开起到关键的作用。目前这三大数据库实行每天进行数据互相交换，使得 3 个数据库所包含的核心数据相同，极大地方便了用户对数据库的使用。

初期的数据库以单纯 DNA 和蛋白质序列为主，每一个数据条目仅包含文件名和序列。但这些数据库大多都已扩展到包含与序列相关的多种信息，包括功能、突变、编码产物、调节因子和参考文献等。除经典的 DNA 和蛋白质序列数据库外，还有生物大分子三维结构数据库（如 PDB）、文献数据库（如 PubMed）、与生物学有关的知识数据库（如 KEGG 和 Gene Ontology）及基因组数据库等多种类型。其中以包含多种数据类型的综合型数据库为今后的发展重点。像 UCSC 的基因组浏览器就是这一类型的很好例子。它集序列、多种基因注解、比较基因组、功能基因组和许多其他数据类型于一体。这类数据库通常具备方便的图形界面，便于不具备生物信息学技能的一般用户使用。但建立这类数据库要求对多种数据类型进行有效的整合，其中不仅需要考虑如何建立数据之间的联系，也对相关的软件技术开发提出新的挑战。

数据格式的建立、数据的准确性和质量控制、方便的数据搜寻方式以及数据的及时更新是数据库建立和维持中的重要问题。目前最为成功和使用最广的序列数据提取系统当首推 NCBI 的 ENTREZ 系统（http://www.ncbi.nlm.nig.gov/Entrez/）。另外，为达到数据库搜寻的最高效率，数据库中数据的重复必须达到最低水平。但要真正做到没有数据重复有较大的难度。NCBI 中的蛋白质序列数据库 nr（non - redunant）就是这类数据库。但事实上这个数据库已经包含相当多的重复序列。另外，还有为数不少的非完整序列的继续存在。而与其相应的 DNA 序列数据库 nt（在 NCBI 的 BLAST 网页上为 nr）已早就公布不再是非重复的了。用户在使用这类数据库时有必要了解这些情况，才能对搜寻结果作出正确判断。

尽管数据库的建立不再是当前生物信息学的焦点，但对现有数据库的扩展、维护仍然非常必要。与此同时，开发新型数据库、研究有效数据格式和类型、促进数据交换和提取也非常关键。有关生物信息学常用数据库的介绍和使用请参考本书第 4 章。

1.3.2 DNA 和蛋白质的序列分析

在现有的序列数据库中找出与用户序列相同或相近序列，可以提供与此序列特征和功能相关的重要信息，涉及的问题是 DNA 和蛋白质序列相似性的分析。其中蛋白质的相似性分析因涉及不同氨基酸的结构和功能的影响不同而远比 DNA 的序列分析来得复杂。序列相似性分析是生物信息学最早所涉及的问题，也是现今生物信息学研究中的日常工作之一。由于数据库的数据量日益增长，其中主要的计算问题是找到一种快速而灵敏的计算机运算法则。可以说以 Needleman - Wunsch 和 Smith - Waterman 为代表的运算法则较为满意地解决了这一问题，但对于新的运算法则的发展仍然十分迫切，以满足数据库快速增长和大规模序列分析所提出的新要求。NCBI 的 BLAST 是现今序列相似性分析中使用最广的常用工具。除此之外，还有最近由 David Haussler 和 Jim Kent 领导的 UCSC 研究组所创立的 BLAT 方法，以极其快速的优点而在其基因组浏览器中得到成功的应用。有关 DNA 和蛋白质序列

分析的详细内容请参考本书第5章。

1.3.3 基因结构的预测

预测一段DNA序列或一个物种的基因组序列中具体哪些区域代表用于编码蛋白质的功能基因是生物信息学研究中的另一经典问题。相对而言，原核生物基因的预测较为简单，因为原核基因没有内含子，因而只需寻找达到一定长度而具有起始密码的开放阅读框(open reading frame，ORF)。其中较为复杂的情形包括：同一转录子编码多个蛋白；不同的基因间相同方向或相反方向互相重叠等。尽管如此，可以说，我们对于原核基因组中的基因预测，已经获得较为满意的成就。然而，真核生物基因因为有内含子和外显子之分，加之选择性转录本的存在，其基因结构的预测成为生物信息学研究中的一大挑战。其中的难点之一为外显子和内含子交界位点的确定，而第一个包含启始编码子的外显子的预测难度为最大。现有的预测方法通常借助于已知蛋白质序列的比对(DNA－蛋白质序列)，与已知cDNA及表达序列标签(EST)的比对(DNA－DNA)，及相近物种基因组序列间的比对(在翻译上的DNA－DNA比对)。除此之外，预测方法利用已知基因结构序列进行训练，采用包括神经网络(neural network)和HMM(Hidden Markov Model)在内的机器学习语言来识别外显子，尤其是外显子与内含子交界区域序列模式特征，已经取得了长足的进展，产生了包括FGENES、GeneFinder、GeneMark、GeneParser、GenScan及FirstEF等在内的一系列预测真核生物基因的软件工具。然而，所有这些称为*ab initio*(意为从头开始，预测不直接依赖实验和其他数据)的预测方法，其准确率均不超过75%，而对第一个外显子的预测准确率最低。真核生物基因预测的另一难点是许多基因利用多种可能的外显子进行不同的组合获得不同的基因表达产物。这是真核生物利用有限的基因组产生复杂的基因功能，达到适应不同发育、生理及环境条件的一种有效途径。我们目前对于什么因素控制或操纵不同外显子的选择仍然知之甚少。很显然，这一研究领域中可提高和完善的空间还很大。这不仅有待于发展更智能化的运算法则，也依赖于我们对于包括RNA剪接在内的许多真核生物基本现象的进一步认识。

1.3.4 蛋白质结构和功能的预测

到目前为止，尽管已知的蛋白质序列已过百万计，但结构已知的仍为少数。到写稿为止，储存在蛋白结构数据库PDB中的条目不过2万左右。因此结构已知的蛋白质数目仍只占已知序列蛋白质总数的2%不到。其原因在于测定蛋白质的结构需要使用非常费时和昂贵的特殊方法，如X－晶体衍射或核磁共振(nuclear magnetic resonance，NMR)。很显然，仅仅依靠实验的方法很难测定所有已知蛋白质的结构。因而，根据蛋白质的一级结构来预测其高级结构，包括预测蛋白质间的相互作用，蛋白质与受体和药物的作用等，具有很大的应用价值。蛋白质在被合成时为线状氨基酸肽链，然后链的不同区域形成包括螺旋(α－helix)、片层(β－sheet)和转折(turn/loop)在内的二级结构。二级结构再互相重叠形成高级结构。蛋白质折叠(folding)的结果一般是把疏水的氨基酸放置在折叠后的内部区域，而把与水和其他分子相作用的极性氨基酸留在表面。一级结构引导折叠的过程，其中有时需要伴护蛋白(chaperone)的帮助。我们目前已经积累了相当数量的已知蛋白质结构，因而可以采用计算的方法来找到或预测哪些序列与已知结构具有相似的结构团(structural fold，指的

是通过相似的折叠连接的相同二级结构方式)。对已知蛋白结构的统计分析表明,蛋白质结构团只有有限的约500种不同的类型,而现有的已知结构已经包括自然界所存在的蛋白结构的90%以上。其中的原因可能是蛋白质折叠过程中化学的限制或是蛋白质结构存在单一的进化途径。由于序列不同的蛋白质可以具有相同的结构,这使得蛋白质结构的预测难度增加。

新型的蛋白结构预测方法采用复杂的统计和机器学习方法,其准确率在逐步地提高。现有的预测方法对于较小的分子(<300氨基酸)的结构预测已经有相当高的成功率,但对于包含多个结构域的大分子的预测则仍然相当困难。这一领域新近的进展包括根据一级结构把蛋白质分成不同的家族,然后用统计的方法来找到每一个家族特定的共同模式(consensus pattern)。本书将在第6章深入讨论结构分析和相关数据库。

1.3.5 基因组数据的分析

由于过去10年内大规模工业化自动基因测序技术的发展和不断成熟,加之全基因组鸟枪法(whole genome shotgun)在基因测序上的成功应用,使得基因组的测序速度得到迅速提高。以目前最先进的设备和技术,一个细菌基因组测序的全部过程可以在一周内完成,而一个平均大小的高等真核生物基因组也可在1~2年内完成。相信在不远的将来,随着新技术的诞生,其中包括纳米技术的应用,基因组测序的速度将进一步加快,成本将成倍下降。很有可能会实现最近有人提出的千元人类个人基因组测序的设想。大量基因组序列的测定,对生物信息学提出新的挑战。其中包括如何提高基因组组装的效率和准确性,如何有效地储存、显示基因组的数据和相关信息,以及如何发展新的软件工具、新的运算法则来比较大量的基因组数据。有关基因组测序、组装和分析的过程请参考本书第3章。

1.3.6 比较基因组和系统发生遗传学分析

大量基因组序列的完成,给我们提供了空前多的DNA数据。对于不同物种基因组的比较分析称为比较基因组学(comparative genomics)。它是基因组学领域里最强有力的、也最具挑战性的研究方法。其研究的着眼点包括:序列的保守与差异、基因组的结构、基因与基因间物理位置的保守、非编码DNA的数量和种类差异等。通过将人类基因组与小鼠及其他模式动物基因组的比较,极大地增进了我们对人类基因组中功能基因的了解。通过比较致病菌和同一物种非致病菌的基因组,可以了解其致病的机制,便于发展控制致病细菌传染的新方法、新疫苗和新药物。对于同一物种中不同个体基因组的比较,可以全面地观测群体中的个体差异(polsmorphism)或多样性。利用基因组数据来研究物种间的进化关系和系统发生,可以克服使用单个基因所存在的片面性,能够从整个基因组的水平上来更为全面地理解物种的进化关系及基因组演变的规律。因而,比较基因组学将是生物信息学中今后的发展重点。与单个或多个基因的序列比较不同,基因组间因所涉及的信息量常常是单基因信息量的上千倍、甚至上万倍,而对所用工具和计算机的内存和速度都提出更高的要求。本书第3章和第8章将分别阐述有关比较基因组学和进化遗传学的相关知识。

1.3.7 功能基因组和蛋白组学数据的分析

基因组的完成给基因表达的研究带来了最为深远的革命。后基因组时代的到来使得生物学的研究重点从以前的以 DNA 测序为主转移到以系统了解基因组内所有基因的生物学功能即功能基因组(functional genomics)为中心。以 DNA 芯片技术为代表的新技术能把一个基因组中所有的基因安放在一张小小的玻璃片上,因而,使得我们可以同时研究所有基因的表达,从而从整个基因组的水平上来研究基因的表达。从 DNA 芯片的设计,信号的定量分析,到根据基因的表达谱来进行分类和组合,找到与表型相关的基因,都包含统计生物信息学方法的应用。更为重要的是根据基因芯片的分析结果,将其与已知的基因和所有相关及相似的多种数据进行整合,对涉及的基因、基因调节、信号和代谢途径进行预测和模型研究。有关基因芯片数据分析的详细介绍请阅读本书第 7 章和第 10 章。

蛋白组学(proteomics)以一个细胞或一个物种所具有的全部蛋白质为研究对象,尽管目前由于技术的限制,无法进行蛋白质的大规模测序分析,其研究的广度和深度仍然有限。但蛋白组代表基因组表达的最后结果和效果,与基因和细胞功能更加直接相关。因此,将会是今后发展的重点之一。其中主要的生物信息学问题是如何有效地根据蛋白质的分子质量和等电点等物理化学特征,以及蛋白质质谱分析、多肽指纹分析(peptide mass fingerprinting)等技术获得的结果,结合已知的基因组数据来预测每一蛋白分离样品的实际身份。

1.3.8 信号传导、代谢和基因调节途径的构建与描述

细胞内的基因和蛋白质都不是独立行使功能。蛋白间互相作用,基因间互相调节,形成一个由信号传导、蛋白网络、代谢途径及细胞间的互相作用等组成的极其复杂和微妙精巧的网络来完成一个细胞的生物功能。过去,我们对生物现象的认识和了解大多仍局限于单个基因(蛋白)或单个信号传导或代谢途径范畴上一种静态的理解。基因组时代和后基因组时代的到来,不仅让我们清楚地认识到从系统和整体水平上来理解细胞功能的重要性,同时也第一次使得在这一水平上的研究成为可能。生物信息学今后的一大任务就是根据比较基因组学、功能基因组学、蛋白组学研究的结果,结合我们从实验生物学积累的所有生物学数据来构建完整的代谢途径及与基因调节和各类信号传导相关的网络系统,包括了解它们之间的相互作用。更具挑战性的是如何有效地表示和利用这类研究的结果和对生物现象在整体和系统水平上进行模拟研究,提出新的理论和学说。

1.4 生物信息学今后的发展方向和趋势

人类基因组计划的完成标志着基因组时代进入高潮和后基因组(post - genome)时代的到来。到目前为止,上千种病毒基因组和近百种细菌基因组及数十种真核生物基因组的测序已完成,更多的物种被列入基因组测序的计划之中。更多详细的有关人类基因组测序工程计划过程和包括其他物种在内的基因组测序进程请参考本书第 3 章。与基因组生物信息学(genome informatics)不同,后基因组生物信息学以从基因组信息获取的生物学知识来了

解生命的基本原则，同时也有其在生物医学应用中的实际目的。后基因组生物信息学与功能基因组中以基因芯片和其他高通量(highthroughput)技术为基础的系统实验学相连，但生物信息学无疑将在实验设计和预计中起着更为主导的作用。

大量生物物种基因组序列的完成和分析及生物信息学研究所带来的新型研究手段和成果正在迅速地改变生物医学的研究方法。其中最显著的一个革命性的改变就是反向遗传学(reverse genetics)研究策略的诞生和大量使用。经典的遗传学研究是从特定的生物表型开始，通过包括定量性状位点(QTL)和细胞遗传学在内的多种实验方法来对相关的基因进行粗略的定位，然后再通过分子生物学的方法，如定位克隆(positional cloning)、序贯缺失(serial deletion)等方法，从数十乃至上百个候选基因中逐一验证，确定具体哪一个基因与所观测的表型相关。这类传统的研究方法周期长，花费大，而且依赖于能观测到的表型，因此很难了解基因组中所有基因的总数及其功能。物种基因组的完成从理论上把一个物种所有的基因都呈现在分子生物学家面前，从而可以采取所谓的反向遗传学的研究策略，即从基因到表型的途径，直接对其生物功能进行研究。许多情形下基因的功能可以通过生物信息学的研究方法来预测，实验研究的目的在许多情形下只是证实预测的功能。而在其他的情形，生物信息学为实验生物学的设计提供信息和思路，缩小研究的对象。因而，我们对基因功能的研究进程将大大加快。与此同时，生物信息学研究将根据比较基因组学、功能基因组学等分支学科的研究成果，运用大规模高度复杂和智能的数学统计模型，从整体和系统的水平上对生物学的现象提出大量新的理论和学说。而对于这些学说和理论的验证将会把实验生物学推向一个新的高度，使生物学的研究进入到系统生物学(systems biology)的新时代。因此，我们有充分理由相信，生物信息学将在21世纪里继续得到迅速的发展，包括与纳米技术相结合，从而对生物医学产生深远的革命性影响。

1.5 生物信息学家应具备的基本知识和技能

正如前面所述，生物信息学涉及生物学、计算机学、数学、统计学等多门学科，从事生物信息学研究的工作者或生物信息学家(bioinformatician 或 bioinformaticist)可以来自以上任何一个领域而侧重于生物信息学的不同方面。事实上，我们今天正需要具备各种背景知识、才能和研究思路的研究人员，集思广益来共同面对生物信息学给我们带来的这史无前例的挑战。然而作为一个生物信息学的工作者，也必须具备一定的基本知识和技能，具体如下。

(1)具备相当程度的分子生物学背景知识，可以是其中的某一方面，比如生物化学、分子生物学、分子物理学或是分子模型，但起码要能理解分子生物学的中心法则。由于生物信息学归根到底是一门生物学科，它的研究最终目的是回答生物学的问题。要编写一个很好的生物信息学软件工具，在具备必要的计算机技能的同时，也必须理解所要解决的生物学问题。进一步而言，真正有效的生物信息学研究应当是带着生物学的问题去寻求计算生物学的解决方法。本书的第2章为缺少生物学背景的读者介绍分子生物学、遗传学和细胞生物学的基本概念。

(2)必须熟悉生物信息学的基本概念和常用分析工具,如 BLAST、CLUSTAL、HMM 及主要基因组浏览器等的基本原理;常用数据库的结构、类型和基本数据内容;主要生物信息学网站,如 NCBI、EBI 等的内容。应当熟悉和掌握至少一两种主要分子生物学软件包的使用,可以是序列分析或分子模型。虽然如下文将提到的,生物信息学家需要有一定的编程能力,但实际上对于许多常规的数据分析,已有不少现成的软件工具。比如用于序列分析的 GCG 和 VECTOR NTI 软件包。本书不同章节中将较为详细地介绍常用的生物信息学软件。熟悉一种软件的使用将有助于学习其他类似的软件。此外,最好能理解常用生物统计中的概率论原理及能使用一些常用的统计软件,包括从最简单的(如微软的 EXCEL)到复杂的(如 S + 、Sigma Plus、SPSS、SAS 等)。有兴趣和条件的读者可阅读参考文献所列的几本介绍生物信息学基本内容及着重于序列分析的参考书[9-11]。

(3)应当适应在使用命令行的计算机环境下工作。使用 Unix 或 Linux 平台便可提供这一方面的经验。由于最早的大中型计算机都是使用 Unix 操作系统,绝大多数用于计算生物学的软件工具是为 Unix 所写。这些软件通常对计算机硬件有严格的要求,因而无法在个人计算机上使用。事实上,相对在个人计算机上使用的操作平台而言,包括微软公司的视窗和苹果公司的 MAC 操作系统,Unix 对于生物信息学的研究具有许多的优点。其中以其可塑性好、便于自动化、软件发放简单等优点尤为突出。Linux 是近 10 年发展起来的类似于 Unix 的操作系统。可以说它就是源程序开放的 Unix 操作系统,因而它具备 Unix 的所有优点,但同时又对硬件没有特殊的要求,可以在个人机上使用。这样一来,其硬件成本较之昂贵的 Unix 系统而言大大地降低。加之其操作系统和相关的大多数软件工具都属开放源程序,可以免费下载,使得 Linux 在过去的 10 年中得到极大的推广使用。与此同时,把成本低廉的多台计算机构成一个机组(cluster)可以用于平行计算(parallel computing),提供非常强大的计算能力。目前用于复杂生物信息学研究项目的计算机机组包含多达数千台计算机。预计今后 Linux 在生物信息学研究中的使用将会更加普遍。因此,掌握 Linux 的使用对于从事生物信息学的研究者来说实属必要。

(4)应有一种以上计算机语言的编程技能,比如 C/C + + ,PERL 或 PYTHON。尽管大多数生物信息学家的日常工作并不是编写复杂的计算机程序,但由于其所涉及的数据通常来自不同的格式,需要经常互相转换或者变换为特殊的固定格式,或是从大量的数据中提出少量特定的信息。但其数据量之大使得手工操作无法实现,又往往难以找到现成的工具。另外,生物信息学家常常做的是把同一分析方法对于大量类似的数据进行重复分析,或是定期性地执行同一任务。在这些情形下,往往可以通过编写简单的剧本(script)程序便可达到目的。在众多的编程语言中,PERL 由于具有易学、方便使用、特别擅长字符串的处理、属于开放源程序等多项优点而在生物信息学研究中最为流行。适用于各种平台的 PERL 软件均可从http://www.perl.org 网站免费获取。而且由于开放源程序运动的结果,大量在生物信息学中所需的常用 PERL 模块和对象已由非赢利的 BIOPERL 机构免费提供,为生物信息学的研究提供了极大的方便。其网址为:http://bioperl.org 和http://cpan.org。PERL 特别适用于生物数据在不同格式之间的转换、从数据中提出特定信息、序列基序的搜寻等生物信息学中的常规工作。尽管比较大而复杂的软件工具大多使用 C/C + + 编写,PERL 也可用于编写相当复杂程度的软件工具。本章作者极力推荐有志从事生物信息学研究的未来同行能够

通过课堂学习或自学掌握 PERL。对于已有其他语言编程能力者，学习并掌握 PERL 应当是非常容易的。行之有效的学习方法之一是在解决实际问题的应用实践中学习、巩固和加强。读者如果想了解 PERL 在生物信息学研究中一些具体的应用例子可阅读参考文献[12]。除 PERL 之外，掌握用于关系数据库的 SQL（sequential query language）语言，对于数据库的建立和数据提取非常有用。另外，掌握一些基本的 HTML 和 Java 编程技能，可以使你具备自行建立网页的能力。

由于生物信息学应用涉及和渗透到生物医学研究的每一个方面，它的基本研究方法和常用技术无疑会在不久的将来成为所有生物医学研究工作者必备的基本技能。生物信息学研究工作者不仅在生物信息学的基础研究和应用研究中起着主导作用，也肩负着推广和普及生物信息学教育方面的艰巨任务。继信息科技（IT）时代之后，我们正期待着生物信息时代的到来！

（梁 平）

参 考 文 献

1 Watson JH, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 1953, 171: 737

2 Bernstein FC, Koetzle TF, Williams GJ, *et al*. The Protein Data Bank: a computer - based archival file for macromolecular structures. J Mol Biol, 1977, 112: 535

3 Kumar A, Ernst RR, Wuthrich K. A two - dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton - proton cross - relaxation networks in biological macromolecules. Biochem Biophys Res Commun, 1980, 95: 1

4 Burke DT, Carle GF, Olson MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science, 1987, 236: 806

5 Kohara Y, Akiyama K, Isono K. The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell, 1987, 50: 495

6 Adams MD, Kelley JM, Gocayne JD, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 1991, 252: 1 651

7 The Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860

8 Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. Science, 2001, 291: 1 304

9 Baxevanis AD, Quellette BFF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2nd ed. New York: John Wiley & Sons, 2001. 94 ~ 105

10 Lesk AM. Introduction to Bioinformatics. Oxford: Oxford University Press, 2002. 62 ~ 114

11 Krane DE, Raymer ML. Fundamental Concept of Bioinformatics. San Francisco: Benjamin

Cummings. 2002. 5 ~ 20

12 Gibas C, Jambeck P. Developing Bioinformatics Computer Skills. Sebastopol: O'REILLY & Associates Inc. 2001. 10 ~ 88

第2章　生物信息的流向

生物信息的流向，或称遗传信息的流向，简称遗传信息流（genetic information flow），是生命活动的基本规律，也是生物信息学的基本研究对象，即利用数理统计结合计算机应用来分析这种遗传信息流，将相关信息整理成数据库以便保存和检索，或者通过程序设计，提供分析软件或服务性网站，对这些信息进行加工处理或预测分析。本章将重点介绍生物信息在细胞内、细胞间和上下代之间的传递规律，借此了解分子生物学、分子遗传学和细胞生物学的基本原理与概念。遗传信息流还包括遗传信息在物种间的交流，这是比较基因组学和进化遗传学的研究内容，将分别在第3章和第8章作相关介绍。最后，本章将讨论这种生物信息流向的紊乱（基因突变）所导致疾病的相关知识。基因组中遗传信息的变异所引起的多态性将在第9章药物基因组学中进行阐述。

2.1　生物信息在细胞内的流向

细胞（cell）是构成生物体的基本单位。细胞内全部信息的集合即细胞组（cellome），可分成不同水平的子集合，包括基因组（genome）、转录组（trancriptome）、蛋白组（proteome）、代谢组（metabolome）、相互作用组（interactome）、定位组（localizome）等[1]。DNA分子或基因中所携带的遗传信息的流向遵循中心法则（central dogma），即由DNA转录（transcription）生成信使RNA，再翻译（translation）成蛋白质。对于RNA病毒，遗传信息的流向首先通过逆转录（reverse transcription）为互补DNA，再按上述中心法则实现遗传信息由DNA到蛋白质的传递。图2-1为中心法则的示意图。

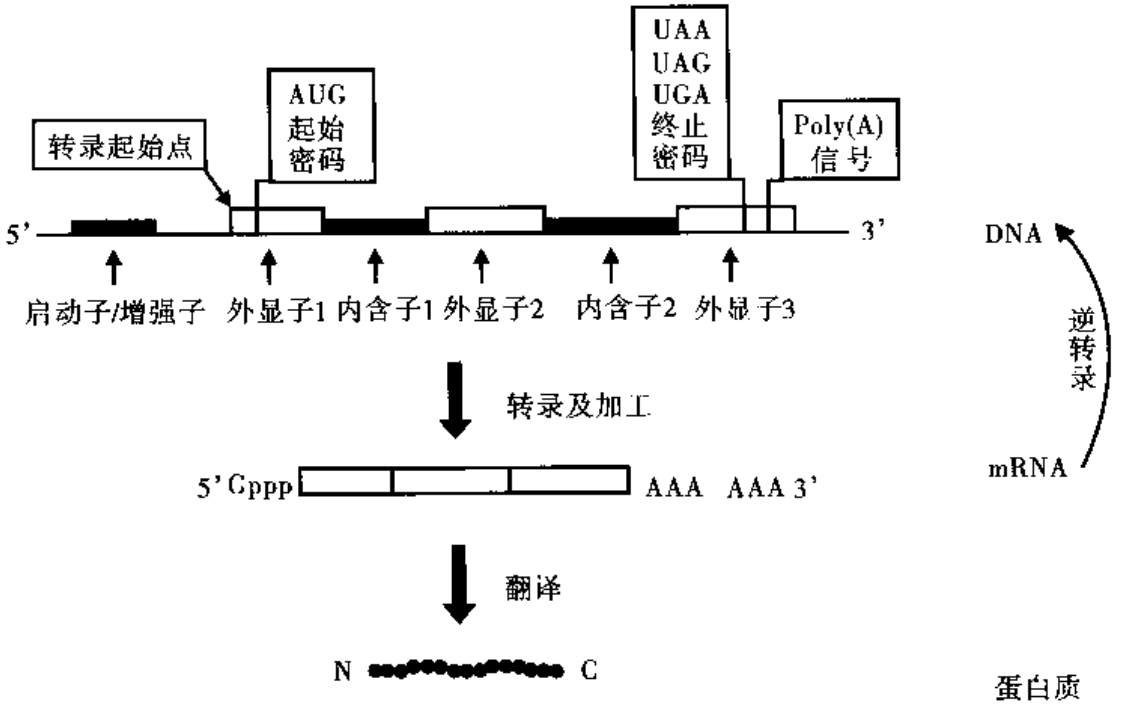


图 2-1 基因的结构及遗传学的中心法则示意

2.1.1 基因组:DNA 和基因的结构

基因组的概念最早在 20 世纪 30 年代提出,是指生物个体中全部遗传信息的总和。

2.1.1.1 DNA 的化学组成与双螺旋结构[2,3]

核酸(nucleic acid)分为脱氧核糖核酸(DNA)和核糖核酸(RNA)两大类,是由嘌呤碱或嘧啶碱(统称碱基)、戊糖和磷酸组成的高分子有机化合物。DNA 分子中的戊糖为 D-脱氧核糖,RNA 分子中的戊糖为 D-核糖。DNA 分子中的碱基含腺嘌呤(A)、鸟嘌呤(G)、胸腺嘧啶(T)和胞嘧啶(C),RNA 分子中的碱基含腺嘌呤(A)、鸟嘌呤(G)、尿嘧啶(U)和胞嘧啶(C)。碱基与戊糖相连形成核苷(nucleoside),后者的戊糖侧链再与磷酸分子结合形成核苷酸(nucleotide)。DNA 或 RNA 分子即是一种多聚核苷酸分子,核苷酸分子彼此以 3',5'-磷酸二酯键相连形成 DNA 或 RNA。分子中保留一个 5'游离磷酸基团(称为 5'端)和 3'游离羟基(称为 3'端)。这就是通常被用来描述核酸分子方向性的 5'→3'或 3'→5'方向。

1953 年,Watson 和 Crick 发现了 DNA 的双螺旋分子结构,这一重大发现被认为是现代分子生物学的重要里程碑。其基本内容是:两股方向相反的 DNA 链相互缠绕形成双螺旋结构(double helix),戊糖和磷酸部分组成该螺旋的骨架并暴露于外侧,碱基朝向内侧,按 A-T,G-C 的碱基配对原则相互以氢键相连。DNA 双螺旋中,每一螺旋的高度为 3.4 nm,由 10 个碱基对组成,每个碱基对沿螺旋轴大约旋转 36°,螺旋的直径为 2 nm。

在实际应用中,常涉及互补链(complementary strand)和反平行链(antiparallel strand)的概念。例如,5'-GATCC-3'的互补 DNA 链为 3'-CTAGG-5',互补 RNA 链为 3'-CUAGG-5',反平行 DNA 链为 5'-GGATC-3'。生物信息学中,为了便于程序编辑,常用表 2-1 中的代码来表示不同类型的核苷酸。

表 2-1 核苷酸代码

核苷酸	代码	核苷酸	代码
A	A	U	U
C	C	C 或 T(U)	Y
G	G	G 或 T(U)	K
T	T	A 或 T(U)	W
非 A	B	A 或 G	R
非 C	D	A 或 C	M
非 G	H	G 或 C	S
非 T	V	A 或 T(U)或 C 或 C	N

2.1.1.2 DNA 的复制、变性、复性与杂交[4]

DNA 的复制是以半保留复制(semi-conservative replication)方式进行,即每条亲链分别作为模板,合成一条互补的子链,这样形成的子代双螺旋与亲代双螺旋完全相同。DNA 复制是从复制叉开始,以 3'→5' DNA 链为模板,按 5'→3'方向延伸。复制起始需要小片段 RNA 为引物,在 DNA 聚合酶作用下,按碱基互补原则依次掺入 4 种脱氧三磷酸核苷酸。DNA 解旋酶、旋转酶、连接酶等也参与 DNA 的复制过程。复制起点可以是多个,同步进行,每个复制单位又被称为复制子(replicon)。

在加热、低盐浓度或碱处理的条件下,DNA 双链解开,碱基间氢键被打断,形成两条单链,此过程被称为变性(denaturation)或熔化(melting)。随着温度升高,双链逐渐打开,最后变成两条单链。这种温度变化范围的中点被称为熔化温度(melting temperature),以 Tm 表示,与 DNA 链中 GC 碱基对含量成正比。一般寡核苷酸片段(15~20 bp)的 Tm 值可通过公式 Tm(℃)=2(A+T)+4(G+C)来粗略计算,已有专用软件(如 Vector NTI)或网站(参见http://www.biotools.idtdna.com/gateway/)计算 Tm 值和模拟熔化曲线。变性后的单链,通过缓慢降低温度,互补的单链又按碱基配对原则重新形成双链,此过程称为复性(renaturation)。基于核酸变性和复性的原理,建立了核酸分子杂交的技术,用于定性或定量检测靶基因,即通常所说的 Southern 印迹和 Northern 印迹。结合 DNA 复制和变性、复性原理,在体外可将靶基因按指数递增方式扩增,称之为聚合酶链式反应(polymerase chain reaction, PCR),这是基因组研究中的革命性技术。

2.1.1.3 基因的结构[2-4]

(1)基因的定义和分类 基因(gene)是 DNA 或 RNA 分子中特定的核苷酸序列,是遗传信息的载体和遗传物质的最小功能单位。对于编码蛋白质的结构基因来说,基因是决定一条多肽链的 DNA 片段。根据其是否具有转录和翻译功能可以把基因分为 3 类。第一类是编码蛋白质的基因,它具有转录和翻译功能,包括编码酶和结构蛋白的结构基因以及编码阻遏蛋白的调节基因;第二类是只有转录功能而没有翻译功能的基因,包括 tRNA 基因和 rRNA 基因;第三类是不转录的基因,它对基因表达起调节控制作用,包括启动基因、操纵基

因和某些假基因。启动基因和操纵基因有时被统称为控制基因。

(2)基因的结构特点　不同基因的大小各异,小的不到 100 碱基对(base pair, bp),大的有几百万 bp。通常一个表达基因含编码区和非编码区。编码区被分隔成不同的区段,称为外显子(exon)。外显子之间的序列称为内含子(intron),属非编码区。内含子序列在 DNA 转录后 mRNA 的加工中被剪切掉,下文"转录组"中将作介绍。不同物种间外显子相对保守,内含子则差别甚大。部分基因仅由一个外显子组成,即不含内含子,转录本序列与基因组序列等同。原核生物和线粒体基因组的基因不含内含子。非编码区还包括 5'和 3'端的序列,通常含转录调控和加工序列,如启动子、增强子、mRNA 多聚 A 加尾信号等,参见图 2-1。多数基因的 5'端含有丰富的未甲基化的 CpG 双核苷酸,称之为 CpG 岛。CpG 富含区通常位于染色体 G 显带中浅着色带。

基因与基因之间的序列被称为无用 DNA(junk DNA),不参与基因的组成,但仍可能对基因的表达发挥调控作用。目前对其生物学功能尚不了解。其组成上除单拷贝外,也含有较多的重复 DNA 序列。重复 DNA 可分为卫星 DNA 和分散重复序列。人类基因组中,单拷贝 DNA、卫星 DNA、分散重复 DNA 分别约占 75%,10%,15%。有关重复 DNA 将在 3.1.3 节中详细介绍。

(3)基因家族和基因簇　多数基因随机分散于不同的染色体上,基因在染色体的位置称为位点(locus)。通常将不同物种间功能相同的、且位于同源部位的基因称为同源基因(homolog 或 ortholog),同一物种间位于不同位点但功能相似的基因称为同系基因(paralog)。理论上,每个基因都隶属于一个基因家族(gene family)。每个基因家族的成员功能相同或相近,可分散于不同染色体或在同一染色体某个区段成簇出现,这种成簇出现的基因又称为基因簇(gene cluster)。如珠蛋白基因簇、免疫球蛋白基因簇、核糖体 RNA 基因簇等。还有一些基因的结构与某些功能基因相似,因进化过程中的突变而不能转录或翻译,失去表达或不具备功能性基因的生物学功能,称为假基因(pseudogene),常用 ψ 表示。如 α 珠蛋白基因家族中存在的一些假 α 珠蛋白基因即 $\psi\alpha$、$\psi\zeta$ 基因。假基因分非加工性(non-processed)和加工性(processed)两大类。前者通常由于基因复制所产生,往往检测不出表达产物;后者是由于 mRNA 经逆转录插入基因组后进一步突变的结果,这种假基因通常不含内含子,但常具有转录甚至翻译功能,表达的 mRNA 或蛋白质不稳定,易于降解,或者所形成的蛋白质无功能。与功能基因相比,假基因往往因无义突变使翻译提前终止,或者突变导致蛋白质的某些结构域缺失而失去功能。近来发现某些假基因也参与基因表达的调控。美国 Yale 大学 Gerstein 实验室已建立相应的人类基因组假基因数据库(http://pseudogene.org)。

基因簇的形成是基因复制(duplication)和演变(divergence)的结果。基因复制可以是一个或几个外显子的复制,也可以是整个或数个基因的复制。整基因的复制是相对常见的,如人类 α-珠蛋白簇和 β-珠蛋白簇的形成就是在漫长的进化过程中,珠蛋白基因不断复制和不等交换(unequal cross-over)的结果,形成了人体不同发育阶段特异表达的珠蛋白基因和假基因。α 类基因簇位于 16 号染色体短臂末端(16p12-pter),该基因簇包括两个功能性的 α 基因(α_1 和 α_2),一个胚胎期表达的 ζ 基因,三个假基因($\psi\zeta_1$、$\psi\alpha_2$、$\psi\alpha_1$)和一个功能未明的 θ 基因,长度约 30 kb。这些基因紧密连锁,其排列顺序沿 5'→3'方向依次为5'-ζ-$\psi\zeta_1$-$\psi\alpha_2$-$\psi\alpha_1$-α_2-α_1-θ_1-3'。β 珠蛋白基因簇位于 11 号染色体短臂(11p15.5),包括 ε、

γ($^{G}\gamma$ 和 $^{A}\gamma$)、δ 和 β 珠蛋白基因，总长度为 70 kb。每条 11 号染色体上只有一个 β 珠蛋白基因，基因簇内各成员也都紧密连锁，其排列顺序沿 5'→3' 方向依次为 5' - $\psi\beta_2$ - ε - $^{G}\gamma$ - $^{A}\gamma$ - $\psi\beta_1$ - δ - β - 3'。在人体发育的不同阶段，各种血红蛋白的合成随着各珠蛋白基因依次表达而呈现相互交替消长。在妊娠 3 ~ 8 周，α 基因簇中的 α 和 ζ 基因与 β 基因簇中的 ε、γ 基因在卵巢囊内表达，相互结合后分别形成胚胎期的 3 种血红蛋白：Hb Gower$_1$($\zeta_2\varepsilon_2$)、Hb Gower$_2$($\alpha_2\varepsilon_2$)和 Hb Porland($\zeta_2\gamma_2$)。在胎儿发育 3 ~ 9 个月，主要合成 α 和 γ 链($\alpha_2\gamma_2$)，ε 和 ζ 链关闭。与此同时，直到出生前，β 基因簇中的 δ 和 β 基因也表达，γ 基因则逐渐减少。至妊娠 36 周，β 链合成迅速增加，至出生后 3 个月，以 α 和 β 链的合成为主，所以胎儿血红蛋白主要为 HbF($\alpha_2\gamma_2$)，初生婴儿占 70% ~ 80%；而成人血红蛋白有 3 种：HbA($\alpha_2\beta_2$，约占 95% 以上)、HbA$_2$($\alpha_2\delta_2$，占 2% ~ 3.5%)和 HbF($\alpha_2\gamma_2$，低于 1.5%)。

2.1.2　转录组：mRNA 的结构[4,5]

2.1.2.1　RNA 的结构

按碱基配对原则将 DNA 分子的遗传信息拷贝到 mRNA 分子中，称为转录(transcription)。转录组是指细胞内经转录生成的全部 RNA 的总和。RNA 因功能的不同，可分为 3 类，即信使 RNA(mRNA)、转运 RNA(tRNA)和核糖体 RNA(rRNA)。除编码 tRNA 和 rRNA 的基因外，所有编码 mRNA 的基因都能翻译成多肽链，最后成为具有一定生物学功能的蛋白质。RNA 分子结构与 DNA 相似，但 RNA 通常以单链形式存在，同一链上不同区域也可按 A - U，G - C 的碱基配对原则形成局部双链次级结构。RNA 的次级结构不像 DNA 的双螺旋结构那样规整，如 tRNA 二级结构呈三叶草型，一些部位形成双螺旋，组成其臂；另一些部位呈环状单链结构。

2.1.2.2　转录过程

转录的过程发生在细胞核内。真核生物中，首先是 RNA 聚合酶识别转录启动子(promoter)。启动子一般位于转录起始点上游 10 ~ 35 bp 内的一段 16 ~ 18 bp 的 DNA 序列，其两端为通用序列(consensus sequence)。转录起始点上游约 10 bp 附近的通用序列为 TATAAT，称为 TATA 盒；转录起始点上游约 35 bp 附近的通用序列为 TTGACA。许多看家基因(house - keeping gene)缺乏 TATA 盒，其转录起始点可为多个，一般位于富含 GC 区。识别启动子的同时，转录激活因子锚定在增强子(enhancer)部位，在转录因子(transcription factor)的作用下，形成转录起始复合物。RNA 聚合酶沿 DNA 模板链 3'→5' 方向，依照 A - U，T - A，G - C，C - G 的碱基配对原则，按 5'→3' 方向合成 mRNA 链，直到转录终止因子识别转录终止子(terminator)，转录过程结束。转录的调节还受沉默子(silencer)的作用，抑制或减弱转录过程。

转录模板链分有意义链、反意义链，前者转录物一般具有编码蛋白质的功能，后者无编码功能，但能调节有意义链的表达。这种反义 RNA(antisense RNA)又被称为干扰 RNA(interference RNA，RNAi)，其发挥调节作用的方式有 3 种：①与 mRNA 结合后，使之不能与核糖体结合，影响翻译的进行；②与 RNA 结合后，影响其稳定性，使之降解，类似 RNA 酶的作用；③与 mRNA 结合发挥类似终止子的作用，使转录提前终止。近年来发现体内双链 *RNA* 在 dicer 的作用下降解为 20mer 左右的双链小分子 RNA，即 siRNA(small interfering RNA)，

可发挥上述反义 RNA 相似的功能。分离和鉴定这类 siRNA 是基因表达调控研究的重要内容。也可通过人工直接合成这种 siRNA 用于封闭某种特定靶 mRNA 的表达，已广泛用于基因的功能研究和基因靶向治疗。将这种 siRNA 导入细胞内的方法一般分两种：一是通过与特殊载体相连经胞吞作用将其运送到胞浆；二是将其导入 DNA 表达载体，转染细胞后进入细胞核，经转录生成 siRNA，再释放到胞浆。进入胞浆内的 siRNA 首先形成双链 RNA 诱导的沉默复合体（dsRNA - induced silencing complex，RISC），RISC 与靶向 mRNA 结合发挥其抑制作用或非特异性 RNA 酶的作用，达到 mRNA 消减（knock - down）的目的。如果导入动物胚胎干细胞，则可复制一些疾病模型。

真核细胞内 RNA 聚合酶分Ⅰ，Ⅱ，Ⅲ类。聚合酶Ⅰ指导 rRNA 前体的合成，最后加工成 5.8S，18S 和 28S rRNA；聚合酶Ⅱ主要负责 mRNA 的合成；聚合酶Ⅲ主要催化小分子 RNA（如 tRNA 和 5S rRNA）的合成。一种称为外显体（exosome）的复合物，参与 RNA 的加工过程，可发挥 3'→5' 核酸外切酶活性，对那些多余的、异常加工的 RNA 进行降解。

2.1.2.3 mRNA 转录后的加工

真核生物成熟 mRNA 是经过转录后加工形成的，该加工过程在细胞核内进行。成熟的 mRNA 才被逸出细胞核，到胞浆中参与核糖体的形成。mRNA 的加工主要包括 5' 加帽（capping）、3' 多聚腺苷酸化（polyadenylation）、RNA 剪切（splicing）和剪辑（editing）等，参见图 2 - 1。新生 RNA 转录本（transcript）达到 25 ~ 30 bp 时，其 5' 端在酶的作用下被共价修饰，添加一个以 5' - 5' 共价键相连的鸟氨酸，以防止细胞内 5'→3' 核酸外切酶对 RNA 的降解。该鸟氨酸的侧链也在酶的作用下进行修饰，其嘌呤环上 N^7 位被甲基化，与此同时，紧接着相连的第一和第二个（仅脊椎动物）核苷酸的戊糖环上的 2' 羟基也被甲基化，完成加帽过程。

新生 mRNA 的 3' 端被酶切下后，再在酶的催化下添加 100 ~ 250 个腺苷酸，此过程称为加尾。这是通过识别多聚腺苷酸化信号 AAUAAA，在其下游 10 ~ 25 个碱基的部位进行切割和加尾。

mRNA 剪切是在一个称为剪切体（spliceosome）的大型复合物参与下进行的，剪切体通常识别内含子两端的 GU - AG 信号，将内含子切除和外显子拼接起来。其他识别序列包括位于内含子 3' 端上游 20 ~ 50 bp 处的 A 以及 3' 剪切点附近的富含 AT 区域。最近发现，一些位于外显子上的外显子剪切增强子（exonic splicing enhancer，ESE）也被认为参与剪切过程。这种 ESE 位点的突变可导致不同的拼接产物的出现。因剪切方式的不同，同一基因最终可产生一种或多种不同的转录本，称为选择性转录本（alternative transcript），每种转录本所含的外显子组成有所不同，特别以 5' 端的不同较常见，这通常由于不同的启动子所引起。选择性转录本的出现是真核生物复杂性的原因之一，即一个基因可有多个转录本，因此可编码多个不同的蛋白异构体（isoform）。如编码 dystrophin 的 DMD 基因，目前至少已发现 7 种不同的分子质量大小不一的异构体，其分布具有组织特异性，功能上也有所不同。

mRNA 的剪辑是在剪辑体（editosome）的作用下进行的，是一种不依赖于 DNA 模板和转录过程的转录后修饰，其结果是导致部分 mRNA 序列的改变，使 mRNA 中出现点突变，包括碱基替换（如 C 经脱氨变成 U）、插入或缺失。目前认为，剪辑仅在人体等少数几种生物中出现，是蛋白质多态性产生的原因之一。

2.1.3 蛋白组:蛋白质的结构、相互作用和定位

基因组和生物信息学研究的核心是理解基因组所编码的全套蛋白质,即蛋白组,或称蛋白质组(proteome)。蛋白组的概念在20世纪60年代已经提出,直到90年代后期才被广泛接受。

2.1.3.1 遗传密码与翻译过程[4,5]

mRNA分子中的遗传信息根据遗传密码(genetic code)规律指导多肽链的合成,称为翻译(translation)。遗传密码是指mRNA中的核苷酸序列所对应的蛋白质序列。遗传密码为一组三联核苷酸(又称密码子,codon),每个三联核苷酸编码一种氨基酸。构成人类蛋白质的20种氨基酸(表2-2)分别由61组三联核苷酸的遗传密码所编码。AUG除编码蛋氨酸外,也是起始密码子,UAG、UGA和UAA为终止密码子,见表2-3。起始密码子与终止密码子之间的所有密码子构成所编码多肽链的开放阅读框(open reading frame,ORF),见图2-1。

表2-2 组成蛋白质的20种氨基酸及其代码

中文名	英文名	三字符代码	单字符代码
丙氨酸	Alanine	Ala	A
精氨酸	Arginine	Arg	R
天冬酰胺	Asparagine	Asn	N
天冬氨酸	Aspartic acid	Asp	D
半胱氨酸	Cysteine	Cys	C
谷氨酰胺	Glutamine	Gln	Q
谷氨酸	Glutamic acid	Glu	E
甘氨酸	Glycine	Gly	G
组氨酸	Histidine	His	H
异亮氨酸	Isoleucine	Ile	I
亮氨酸	Leucine	Leu	L
赖氨酸	Lysine	Lys	K
蛋氨酸	Methionine	Met	M
苯丙氨酸	Phenylalanine	Phe	F
脯氨酸	Proline	Pro	P
丝氨酸	Serine	Ser	S
苏氨酸	Threnine	Thr	T
色氨酸	Tryptophan	Trp	W
酪氨酸	Tyrosine	Tyr	Y
缬氨酸	Valine	Val	V

表 2－3　遗传密码表

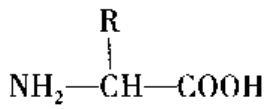
第一碱基	第二碱基 U	第二碱基 C	第二碱基 A	第二碱基 G	第三碱基
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	终止密码	终止密码	A
	Leu	Ser	终止密码	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	起始密码				
	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

蛋白质的翻译过程是在核糖体中进行的。核糖体位于粗面内质网上或游离于胞浆中，分大、小亚基两部分。原核生物和真核生物核糖体组成有所不同，真核生物核糖体小亚基(40S)主要由 18S RNA 组成，大亚基(60S)则由 5S，5.8S 和 28S RNA 组成。翻译过程较复杂，涉及上百种生物大分子的参与。翻译由起始密码 AUG 开始，沿 mRNA 的开放阅读框 5'→3'方向进行。首先在起始因子和 GTP 的作用下，识别 mRNA 5'帽和 3'多聚 A 尾，携带起始密码(5'AUG)的反密码子(anticodon)（5'CAU)的蛋氨基酰－tRNA 与结合在核糖体小亚基上的 mRNA 中对应的密码子配对结合，形成起始复合物。起始密码 AUG 通常位于 5'帽下游 100 bp 附近，起始密码常含有独特的 Kozak 识别序列 GCCRCCAUGG(R 代表 A 或 G)。起始复合物进一步与核糖体大亚基结合，释放起始因子和 GTP。多肽链的延伸是在延伸因子、转肽酶等的作用下，经反复的进位、转肽和移位而进行的，直到出现终止密码，肽链合成结束。mRNA 经去帽和去尾后被降解。蛋白质翻译后，一般须经过进一步修饰，如糖基

化等，才形成有功能的蛋白质。

2.1.3.2 蛋白质的一级结构[5,6]

蛋白质是由20种不同氨基酸组成的多肽链。氨基酸的基本结构为：



各种氨基酸之间的区别在于侧链基团R的不同。通常根据R基团将20种氨基酸分为4类，即极性、非极性、酸性和碱性氨基酸。氨基端称为N—末端，羧基端称为C—末端。氨基酸之间通过肽键(—CO—NH—)相连形成多肽链，保留N—末端的氨基和C—末端的羧基。多肽链中氨基酸的排列顺序，称为蛋白质的一级结构(primary structure)。

2.1.3.3 蛋白质的二级结构和三维构象[5,6]

蛋白质一级结构的肽键中，氧原子带部分负电荷，氢原子带部分正电荷。由于多肽链中含几十、几百，甚至上千个肽键，氢氧原子之间可形成氢键，从而形成α螺旋(α-helix)或β片层(β-sheet)。α螺旋类似线圈(coil)，就像电话机上的电话线圈一样，每螺旋一圈含3.6个氨基酸残基；β片层正如起伏相间的板层相叠结构。另外，邻近的不同半胱氨酸的巯基间可形成二硫键(—S—S—)，也是导致蛋白质折叠的重要原因，可形成环状结构(loop)。多肽链α螺旋、β片层和二硫键的形成便是蛋白质的二级结构(secondary structure)。二级结构之间还相互折叠形成β-β片层结构、α-α螺旋、α螺旋-β片层、环-环结构等。线圈绕线圈(coiled-coil)是一种特殊的超二级结构，相邻的两个α螺旋间相互包绕，形成疏水键相互作用，侧链间还可形成离子键，这种特殊的结构方式是由蛋白质一级结构所决定，通常是由于七氨基酸重复单位(heptad repeat)组成，可通过软件作预测分析。线圈绕线圈结构通常是复合体中不同亚基相互作用的区域。

由于多肽链的组成不同，极性或酸、碱性氨基酸，具水溶性，称为亲水性(hydrophilic)；非极性氨基酸，难溶于水，称为疏水性(hydrophobic)。对于所形成的球状多肽链而言，亲水性基团常位于表面，疏水性基团一般位于内部。对于膜内在蛋白质而言，亲水基团朝向膜的外侧，疏水基团朝向膜的内部，这种蛋白质与周围环境之间的亲水和疏水作用便形成了蛋白质的三级结构(tertiary structure)。每种蛋白质都有其独特的三级结构。具有三级结构的蛋白质称为单体(monomer)。一般只有具有三级结构的蛋白质才能发挥其相应的生物学功能。

许多活性蛋白质不是由一个单体组成，而是由多个相同或不同的单体相互作用形成的多聚体，每个单体又称为亚基(subunit)。这种结构称为蛋白质的四级结构(quaternary structure)。如血红蛋白是由2个α和2个β亚基组成的四聚体。每个亚基不能发挥作用，只有当这种四聚体与亚铁结合后，才能发挥携带氧的功能。蛋白质的二、三、四级结构统称为蛋白质的空间构象或三维构象(3D conformation)。在加热、酸、碱、盐等条件下，蛋白质的空间构象破坏，称为蛋白质的变性作用；如果变性条件温和，蛋白质又可恢复其空间构象，称为蛋白质的复性。

目前已有不少蛋白质分析软件用于预测蛋白质的二级结构、结构域、结合部位和相互作用模型。这些软件是在蛋白质一级结构基础上，通过氢键和疏水键形成、化学自由能分析等进行预测，其准确性据称可达70%左右，但假阳性和假阴性率也高达30%～40%，详见第6章有关结构生物信息学内容。这些预测有待于实验的进一步校验。蛋白质结构分析中的两大关键技术是X－衍射晶体分析(X－ray crystallography)和核磁共振技术(NMR)。目前美国NIH已成立蛋白质结构分析协作组，以完成对重要蛋白质结构的实验分析，并将所有的蛋白分子结构收录于PDB数据库中(http://www.pdb.org)，这是蛋白组学和结构基因组学的重要组成部分，详见第6章。此外，要了解一个蛋白质的功能，除了解其本身的结构特点外，分子进化(molecular evolution)分析也有助于理解蛋白质的功能，参见第8章有关进化遗传学内容。

2.1.3.4 蛋白质的相互作用和定位[4-6]

(1)蛋白质的相互作用 蛋白质的一级结构决定其空间结构，蛋白质的结构又决定其功能。许多蛋白质在发挥其生物学功能时，往往某个或某些氨基酸是其关键区域，称为结构域(domain)。功能相同或相近的同源蛋白质常具有相同的结构域，每个结构域相对保守，有其共同祖先，是由于进化过程中的复制、演化、转位等形成的。多数蛋白质发挥作用时是与其他蛋白质协同作用，不同蛋白质之间形成复合体(complex)。每个蛋白质可视做这一复合体的亚基(subunit)，亚基间相互作用，形成紧密的复合体结构或共同组成复合体的活性中心。这种相互作用往往涉及蛋白质的结构域，当结构域中的氨基酸因突变被替换后，会导致复合体去稳定化(destablization)，复合体形成障碍而被降解。如下文"溶酶体相关细胞器"形成中所涉及的BLOC－1复合体中，dysbindin，pallidin和muted蛋白组成复合体的亚基，由于基因突变导致dysbindin缺乏后，pallidin和muted也因去稳定化而被降解[12]。

抗原－抗体结合也是一类特殊的蛋白间相互作用，通常在制备抗体前，需要根据与其相互作用的抗原决定簇(epitope)来合成或设计抗原，已有专用软件分析特定蛋白质的抗原决定簇。蛋白质的相互作用可以通过生物信息学方法加以预测(详见第6章)，但目前主要还是通过实验的方法加以确认蛋白间的相互作用。这些方法包括酵母双杂交系统(yeast two hybrid，Y2H)、免疫共沉淀(co－immunoprecipitation，co－IP)、GST标记蛋白共分离(GST－pull down)技术等。

广义的蛋白相互作用除形成复合体外，也包括同一生化代谢途径或信号传导途径的相关联的一系列酶或蛋白质，详见2.1.4"代谢组"的内容。

(2)蛋白质的定位 膜结构和细胞骨架(cytoskeleton)是真核细胞的两种最基本的组成结构。膜结构包括质膜、细胞器膜、核膜，其功能主要是将特定的代谢途径限定在一定场所内或膜表面进行。细胞浆内的生化反应以及细胞器的分布则是通过与细胞骨架的相互作用而调节。蛋白质的定位可粗略分为膜关联蛋白和非膜关联蛋白两大类。也就是说，细胞内生化反应体系的酶和蛋白质可定位于细胞浆、细胞核、膜内部、膜表面、细胞器内或与细胞骨架相关联。明确每种蛋白质的细胞内定位有助于了解其生物学功能和病理改变。

胞浆蛋白又称可溶性蛋白，不与任何细胞器相关联，通常在胞浆核糖体内合成，作用于胞浆内可溶性代谢物质，或与细胞骨架相互作用。胞浆内的游离核糖体合成的蛋白质也被直接运送到细胞核内或线粒体、过氧化物体中。核蛋白的特异性定位由其内部的7～9个呈

碱性的氨基酸(又称细胞核定位序列)所决定,在胞浆内合成后,通过核膜小孔进入细胞核内。线粒体和过氧化物体的特异性定位分别由 N 端 12 ~ 30 个两性氨基酸和 C 端 SKL 序列所决定。其余细胞器的蛋白是由内质网合成并经 Golgi 体修饰后,通过小泡转运而得到,其转运方式参见 2.1.5.5“小泡转运”。这类细胞器蛋白前体(preprotein)的 N 端通常具有一段前导序列(leader sequence)或称信号肽(signal peptide),引导蛋白进入内质网后,被蛋白酶分解。分泌型蛋白也是通过小泡转运方式运送到细胞外的。

蛋白质转运前须经过折叠(folding),一般在内质网中进行。先经二硫键异构酶作用形成二硫键,进一步受多肽异构酶作用形成空间构象,再在伴护蛋白(chaperone)的协同下进一步折叠和自我装配。蛋白质折叠后可防止其渗漏出内质网或被降解。只有按正确方式折叠的蛋白质才能被小泡运送出内质网,进入 Golgi 体。未按正确方式折叠的蛋白被泛素(ubiquitin)结合,在胞浆中的蛋白体(proteosome)内被蛋白酶降解,参见图 2 -2 和图 2 -3。

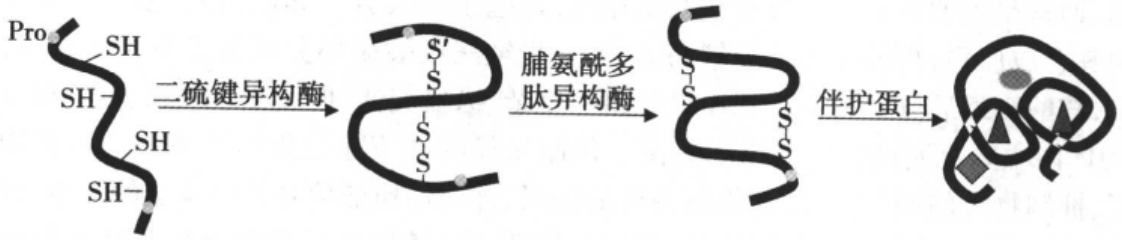


图 2 -2　蛋白质在内质网中的折叠

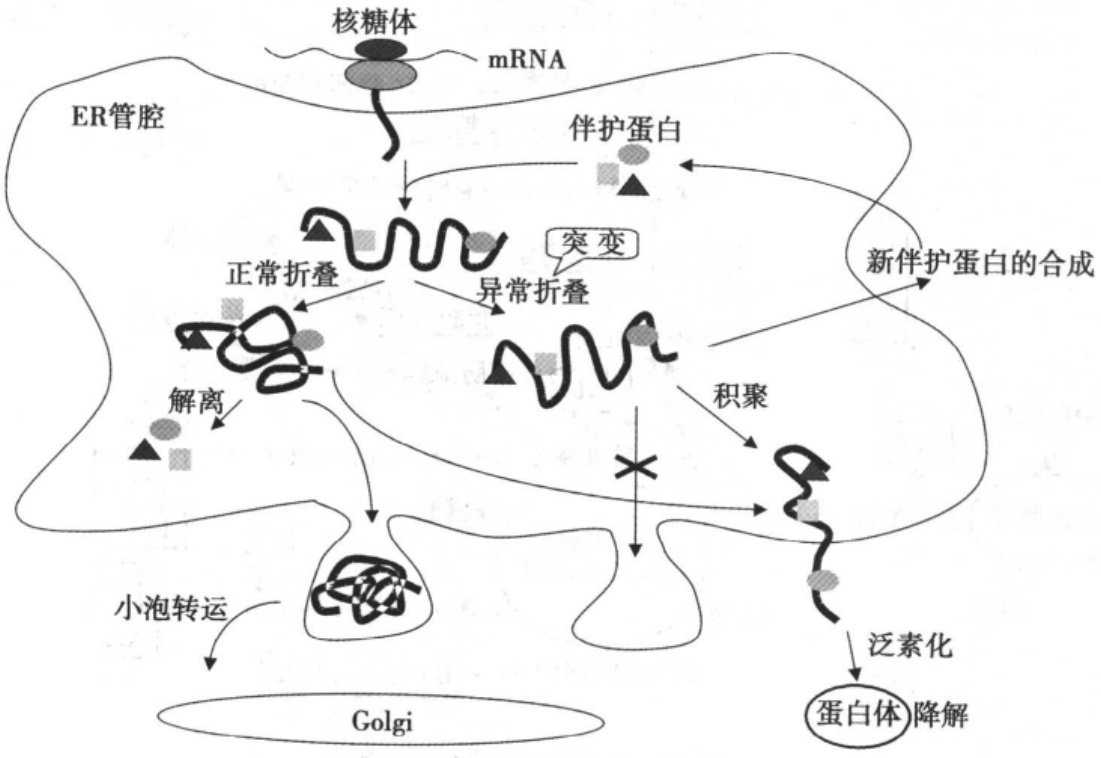


图 2 -3　蛋白质折叠与降解示意图

研究蛋白质的定位一般需要借助特殊的标记或分离技术，通常需要用已知的蛋白作为对照，进行共同定位。常用的方法如免疫荧光技术、激光共聚焦、免疫电镜和扫描电镜技术、细胞器分离/质谱分析技术等。蛋白质的定位还与其一级结构有关。通过生物信息学分析蛋白质的结构域或基序(motif)，也可对某些新发现蛋白质的定位进行预测。

2.1.4 代谢组：生化代谢途径和信号传导途径[4~7]

2.1.4.1 生化代谢途径

人体内的化学反应通常是由多种反应有机地结合并按特定方式进行的，称为代谢途径(pathway)。体内代谢(metabolism)分物质代谢和能量代谢两种，按代谢物的去向又分为合成(anabolic)代谢和分解(catabolic)代谢两种途径。合成代谢是指合成原料氨基酸、单糖、脂肪酸、含氮碱基等在一系列酶促反应下合成生物大分子(如蛋白质、多糖、脂类、核酸等)的过程，通常是需要能量的。分解代谢则相反，这些生物大分子如蛋白质、多糖和脂类先被降解为其合成原料，再转变为乙酰辅酶A，进一步经三羧酸循环分解为二氧化碳和水，并释放能量，核酸则最后分解为氨。代谢过程错综复杂、纵横交错，但又相互联系、相互调节。图2-4为体内主要代谢途径的网络示意图。代谢途径的调节除受化学平衡反应的底物或产物浓度、抑制物等的调节外，还受细胞内外的激素、神经递质的信号传导系统的调节，使细胞内的代谢遵循一定的规律，以满足机体活动的需要。已建立有关代谢网络及其调节的专门数据库KEGG(http://www.genome.ad.jp/kegg/)。

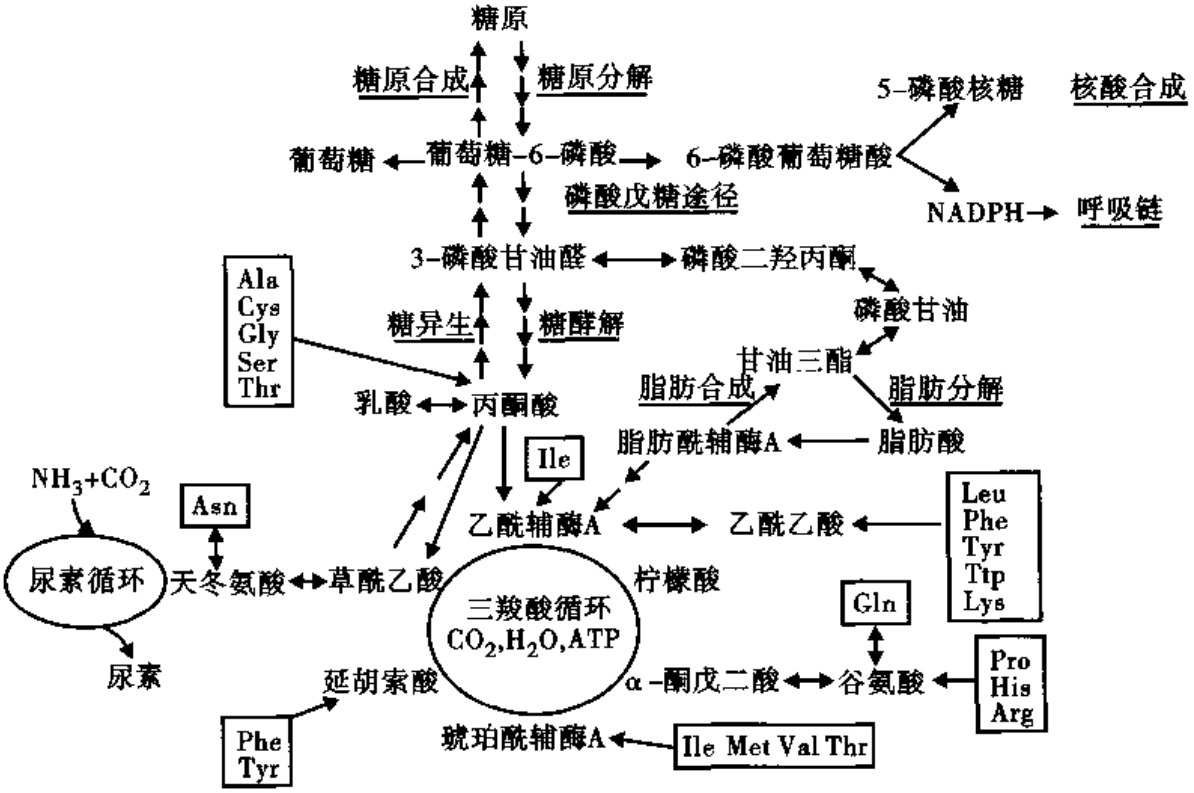


图2-4 生化代谢途径的相互联系

2.1.4.2　信号传导途径(signal transduction pathway)

细胞与细胞间的交流与对话是根据细胞对外界信号的反应而进行的。这种细胞间的交流可以依靠细胞的相互接触或缝隙连接,但更远的交流只有依靠体液中的激素或神经递质的信号传导系统。信号传导到细胞内一般是通过细胞表面受体或细胞内受体介导。一些类固醇激素、维生素D等脂溶性物质可直接透过质膜与位于胞浆或细胞核内的特异性受体结合,形成受体复合物,与DNA的调节部位结合,影响基因的表达。激素或神经递质的信号传导更主要的方式是通过细胞膜表面受体结合,引起构象的变化,产生跨膜信号,触发细胞内一系列级联反应(cascade),最终影响基因的表达。这些受体传导途径主要分为3类:有些是与膜离子通道相互作用;有些是膜受体本身具有酪氨酸蛋白激酶的活性,对靶蛋白中Tyr残基进行磷酸化;有些是第一信使与膜受体结合后激活第二信使而发挥作用,参见图2-5。常见的第二信使包括环核苷酸、钙离子-磷脂酰肌醇、一氧化氮(NO)、鞘氨醇(ceremide)等。参与信号传导途径的蛋白质种类繁多,主要包括蛋白激酶(protein kinase)、蛋白磷酸酶(protein phosphatase)、鸟苷三磷酸酶(GTPase)、受体(receptor)、衔接蛋白(adapter protein)等。近年来发现多聚磷酸也在信号传导中发挥重要生理作用。

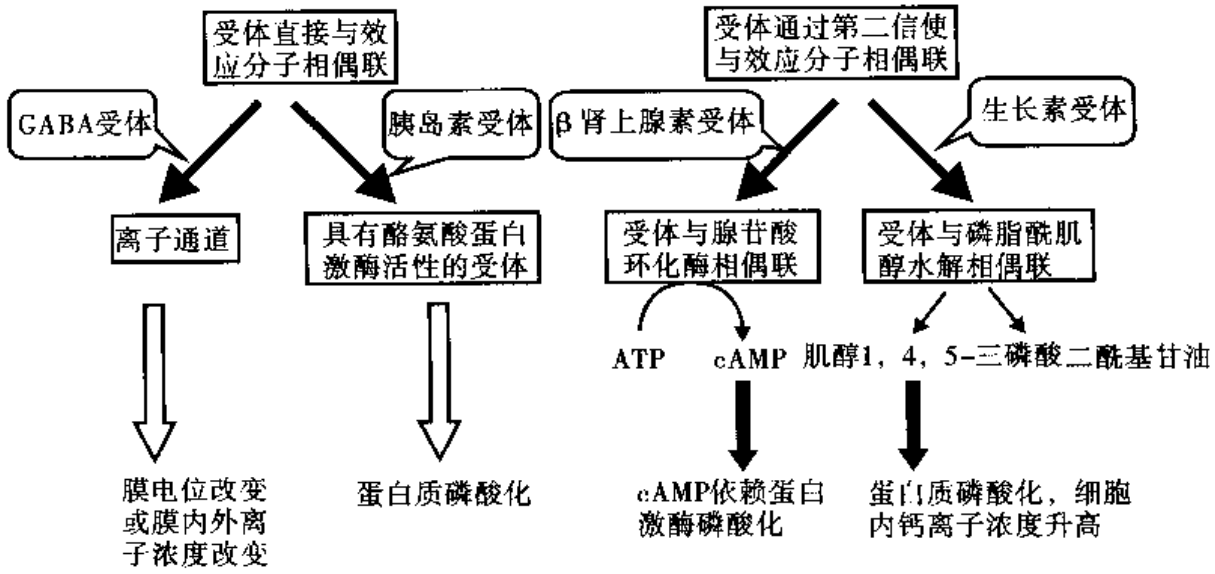


图2-5　信号传导中受体的作用

(1)激酶组　蛋白质磷酸化(phosphorylation)是蛋白质翻译后修饰中的最常见方式,可发生在一系列代谢过程或信号传导途径,对体内代谢酶活性的调节、信号传导的级联反应、细胞运动、膜通道活性、细胞核的装配、细胞周期调节等都起着非常重要的作用。蛋白质磷酸化后可影响其与其他蛋白或配体(ligand)的相互作用,或引起其构象变化,导致其活性改变,或者产生新的结合位点。磷酸化的结果可以使某一途径开放,也可使之关闭。体内磷酸化过程都是在一类称为蛋白激酶(总称激酶组[8],kinome)的作用下进行的,其逆过程即磷酸基团的水解,则是在蛋白磷酸酶的作用下进行的。

真核生物中,99%以上的蛋白质磷酸化发生在丝氨酸(Ser)或苏氨酸(Thr)残基上,只有极少数发生在酪氨酸(Tyr)残基上,如上面提到的一些具有酪氨酸蛋白激酶活性的膜受体,

它们对信号传导系统起着重要的调节作用。此外,组氨酸(His)或天冬氨酸(Asp)也可能发生磷酸化,但对信号传导的作用尚不明确。因此,参与信号传导系统的蛋白激酶主要分为两大类,即 Ser/Thr 蛋白激酶和 Tyr 蛋白激酶。这些激酶分属不同的蛋白家族,常具有一定的结构域,参与不同的调节过程,发挥不同的生理功能。

(2)第二信使系统　作为第二信使的环核苷酸主要为环腺苷酸(cAMP)和环鸟苷酸(cGMP)。以 cAMP 的形成为例,是在腺苷酸环化酶的作用下,由 ATP 转变而来,腺苷酸环化酶活性本身又受 G 蛋白的调节,Gs 蛋白促进其活性,Gi 蛋白抑制其活性。G 蛋白是体内广泛存在的一类 GTP 结合蛋白,又称鸟苷酸三磷酸酶,简称 GTPase,主要包括以下几类:小分子 GTPase(如 Arf, Rab, Ras, Rho, Ran, Sar 等),三亚基 G 蛋白,延长因子,转位 GTPase, Dynamin 等。参与调节腺苷酸环化酶活性的 G 蛋白属于三亚基 G 蛋白,其静态形式(失活)与 GDP 结合,当与 GTP 结合后,其他亚基解离,剩余的 α 亚基才具有活性。G 蛋白本身进入再生循环,在一系列酶作用下形成 GTP 或 GDP 结合形式。信号传导途径中,这类 G 蛋白首先与膜七螺旋(seven - helix)受体结合后被激活,促使 cAMP 的生成。体内有一系列 cAMP 依赖的蛋白激酶,这类激酶由催化亚基和调节亚基组成。cAMP 与调节亚基结合后,释放出催化亚基,可催化靶蛋白中 Thr 或 Ser 残基上的羟基磷酸化,磷酸化蛋白则发挥其相应的生理作用。添加的磷酸基团可在磷酸酯酶作用下去磷酸化,cAMP 也在磷酸二酯酶的作用下水解为 5' - AMP 而失去其第二信使的功能。cGMP 的形成和作用与此相似,其不同点是鸟苷酸环化酶属于受体的一部分,结构类似酪氨酸蛋白激酶。

许多受体与激素或神经递质结合后,激活与膜结合的磷脂酶 C,其激活方式与腺苷酸环化酶相似,需要 G 蛋白参与。活化的磷脂酶 C 可将膜结合的磷脂酰肌醇 1,4,5 - 三磷酸酯水解为肌醇 1,4,5 - 三磷酸和二酰基甘油,两者均可作为第二信使发挥作用。肌醇 1,4,5 - 三磷酸与内质网相应受体结合后,可促进钙池中钙离子的释放,使胞浆内钙离子浓度迅速增加,与钙调蛋白(calmodulin)形成复合物,使蛋白构象发生改变,许多蛋白质就是受这种钙调蛋白复合物的调节而发挥作用。肌醇 1,4,5 - 三磷酸可被迅速去磷酸化生成肌醇 1,4 - 二磷酸和肌醇 1 - 磷酸,后两者不能发挥第二信使的功能。二酰基甘油作为第二信使可促进蛋白酶 C 与钙离子的结合而使该酶被激活,参与一些蛋白质磷酸化而发挥作用。二酰基甘油和肌醇 1,4,5 - 三磷酸常协同作用。

NO 被发现作为血管平滑肌松弛剂,具有扩张血管功能。其作用方式也是作为第二信使发挥作用。NO 的形成是在 NO 合成酶的作用下,经 NADPH, O_2 参与,是由 L - 精氨酸转变为瓜氨酸的反应中生成的。FMN、FAD、血红素以及四氢生物蝶呤是该酶的辅酶。肿瘤坏死因子或某些细胞因子如白介素 - 1,可促进 NO 合成酶的合成。NO 合成后,可活化胞浆内的鸟苷酸环化酶,生成的 cGMP 可能激活蛋白激酶 C,使一些蛋白质磷酸化,导致血管平滑肌松弛。此外,NO 也具有防止血小板凝集,促进炎症反应中巨噬细胞的吞噬功能,还能发挥中枢内神经递质的作用。

(3)受体组:信号传导受体的分类与效应　信号传导系统极其复杂。药物或体内一些物质可作为受体的拮抗剂(agonist)发挥作用,同时信号传导中又存在反馈环,以促进或抑制传导过程。细胞内各类受体的总和称为受体组(receptome)。信号传导系统中涉及的受体以膜受体为主。图 2 - 5 为信号传导受体途径示意图。以下列出一些常见的信号传导过程

的受体途径及其效应。

1)受体/离子通道→膜去极化或膜渗透性改变;

2)七螺旋受体→三亚基 GTP 结合 G 蛋白→多种不同效应;

3)受体/组氨酸激酶→调节物→基因表达、鞭毛运动等;

4)酪氨酸激酶受体→ras, MAP 激酶,PLC,PI3 激酶→细胞增殖、分化等;

5)细胞因子受体→JAK 激酶→STAT 转录因子→基因表达;

6)酪氨酸激酶偶联受体→胞浆酪氨酸激酶→基因表达;

7)受体 - 苏氨酸/丝氨酸激酶→Smad 转录因子→基因表达调控;

8)膜鸟苷酸环化酶受体→cGMP→激酶或通道蛋白的调节→血管扩张、受精等;

9)神经鞘磷脂酶偶联受体→鞘氨醇活化激酶→基因表达、细胞凋亡等;

10)整合素(integrin) →非受体性酪氨酸激酶→多种反应,如细胞运动、基因表达等;

11)选择素(selectin) →细胞黏附;

12)Cadherin→接触抑制;

13)Notch→细胞命运;

14)刺猬(Hedgehog)受体→细胞分化;

15)细胞浆鸟苷酸环化酶受体→cGMP→激酶、通道蛋白等→NO 第二信使系统;

16)细胞浆类固醇受体→激活转录因子→基因表达。

(4)信号传导通路 以上分别叙述了信号传导途径中一些关键的物质。体内信号传导通过整合这些物质,构成一条完整的通路,完成细胞外信号向细胞内的传导。这些传导通路根据信号的种类不同,选择的途径也不同。每条通路又涉及一系列的步骤,极其复杂,有些通路已基本明确,有些通路中的某些环节尚不清楚,新的信号传导方式尚在不断发现之中。目前基本明确的几种主要信号传导方式包括以下几种。

1)G 蛋白偶联、经七螺旋跨膜受体的传导通路:如嗅觉传导、视觉反应、经 β 肾上腺素受体的代谢调节等。

2)影响基因表达的传导通路:如类固醇激素、甲状腺素激活核内受体;带电荷的小分子、多肽或蛋白质(如细胞因子、TGF $-\beta$、TNF、Notch、Hedgehog、Wnt、TCR 等)经膜受体激活胞浆内受体。

3)由分裂素活化蛋白激酶(mitogen - activated protein kinase, MAPK)到细胞核的传导通路:如生长或分化因子(含癌基因和抑癌基因产物)、应激反应、细胞凋亡等。多细胞生物中,已明确 3 种常见的 MAPK 传导途径:①细胞外信号调节激酶(extracellular signal - regulated kinase, ERK)途径,控制细胞分裂,其抑制剂多用做抗癌药物;②c - Jun 氨基末端激酶(c - Jun amino - terminal kinase, JNK)途径,是转录过程的关键调节物,JNK 抑制剂可用做类风湿性关节炎的治疗;③p38 MAPK 途径,被炎症性细胞因子或环境应急因素激活,通常与哮喘和自身免疫疾病有关。MAPK 磷酸化中继系统参见图 2 - 6[9]。

(5)常用的信号传导途径网站[5]

1)三类 MAPK 途径:http://kinase.oci.utoronto.ca/signallingmap.html。

2)信号传导途径概况及其与 PubMed 链接:http://www - personal.umich.edu/%7Eino/List/。

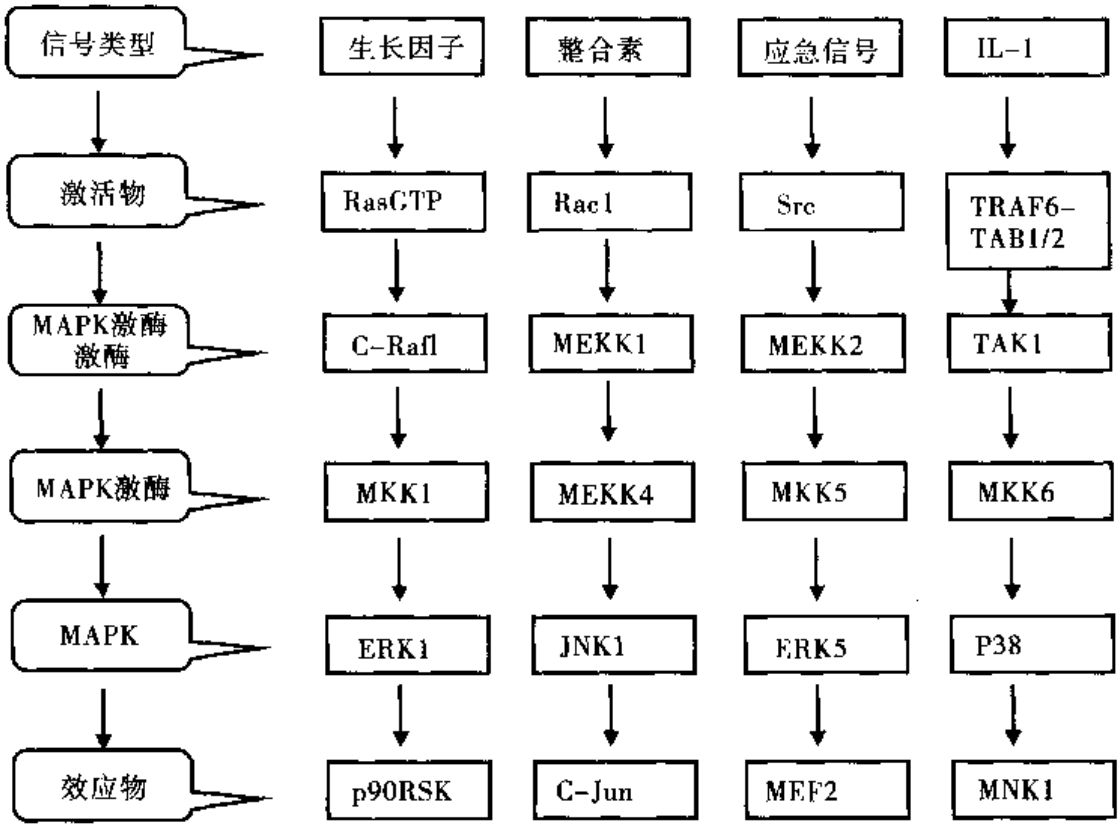


图2-6 信号传导中MAPK磷酸化中继系统

3)信号传导基因序列:http://vlib.org/Science/Cell_Biology/signal_transduction.shtml。

4)信号传导研究常用抗体:http://www.cellsignal.com/。

5)信号传导途径介绍:http://stke.sciencemag.org/。须付STKE使用费,必须是美国科学促进会(AAAS)会员才能进入。

2.1.5 细胞组:细胞器的结构与功能[5-7]

生物体内每种细胞都担负起特定的生物学功能。每个细胞都会经历一系列的活动,如与周围环境的相互作用;合成或再生各自细胞内的成分;对理化因素引起的损伤进行修复;细胞生长和繁殖;产生能量供细胞活动所需。这些基本的细胞活动都是在细胞器内完成。传统意义上所谓的细胞器,是指光学或电子显微镜下可见的细胞内被膜所包绕的具有独特形态特征的结构,如质膜、内质网、Golgi体、溶酶体、过氧化酶体、线粒体等。细胞组是研究细胞内所有细胞器、细胞核、细胞浆以及细胞外基质的结构与功能,细胞间的相互作用等,是细胞生物学的主要研究对象。

2.1.5.1 质膜(plasma membrane, PM)

细胞的膜结构是由脂类和蛋白质共同组成的脂质双层结构。脂类主要为磷脂和胆固醇。磷脂的极性头朝向膜的细胞内或细胞外基质(ECM)侧,疏水的脂肪酸链尾位于内部,形成脂双层(bilayer)结构。胆固醇的羟基与磷脂的极性头相互作用,其余疏水部分排列于脂双层内部。脂双层结构具有流动性和不对称性的特点。膜蛋白分内在蛋白(integral pro-

tein)和外周蛋白(peripheral protein)两类。前者穿透或嵌入脂质双层,后者位于脂质双层的表面,膜脂或膜蛋白常被糖基化。这便是著名的细胞膜流动镶嵌模型。

质膜是细胞膜结构中最具代表性的一种,它将细胞浆(或称细胞质)包绕起来,组成细胞的外屏障。膜脂可形成脂簇(lipid rafts),在细胞内外物质交换中起重要作用,膜脂还可参与信号传导。膜蛋白主要发挥:①受体(receptor)、通道(channel)、载体(carrier)、泵(pump)的运载功能;②进行细胞内外的信号传导;③指导细胞的靶向运动;④介导细胞与细胞或细胞与细胞外基质间的相互作用。

细胞的运载功能和信号传导功能决定细胞内的一系列生化反应的发生和生物学功能的发挥,是细胞最基本的活动形式,也是细胞的生物学功能中最为复杂的形式。

人体内或许多高等动物中含有200多种不同的细胞类型。这些不同细胞类型的相互作用即构成组织和器官。细胞与细胞的相互作用主要通过质膜来完成。细胞与细胞可形成不同形式的连接方式,如屏障连接、黏附连接、缝隙连接、突触连接等。细胞与基质间也形成黏附连接。这些相互作用,使细胞与细胞、细胞与周围环境之间形成交互连接网络,可进行物质与信息的交流。

2.1.5.2 内质网(endoplasmic reticulum, ER)

内质网是许多细胞中最明显的细胞器,占所有膜结构的一半以上。其形态各异,呈层叠或网槽(cisternae)状,或分枝细管状,或扩张的囊状结构。所形成的内腔体积约占整个细胞的10%。ER的主要生物学功能是除线粒体和过氧化物体以外,所有膜结构中脂质和蛋白质合成的主要场所,也是分泌型蛋白质和细胞器中腔内蛋白合成的主要场所,同时ER还参与解毒和钙离子的调节作用。ER又分为粗面内质网(RER)和滑面内质网(SER)两类。粗面内质网是由于其胞浆面含有丰富的核糖体。RER是蛋白质合成的主要场所,SER则是解毒和脂类合成的主要场所。

2.1.5.3 高尔基体(Golgi Apparatus)

高尔基体是由扁平网槽状膜结构组成。高尔基体具有方向性,接收货物的一面称为顺面(*cis*),输出货物的一面称为反面(*trans*)。高尔基体被比做细胞的邮局,对来自ER的货物(蛋白质和脂类)进行包装(小泡形成)和修饰(如糖基化),然后将货物集中到货运站(反式高尔基网络, trans-Golgi network, TGN),再按指定地点将货物发送出去。货物去向主要为质膜、出胞或到溶酶体。

2.1.5.4 溶酶体(lysosome, LYS)

溶酶体被喻做细胞的消化器官。由膜结构包绕,内含丰富的酸性水解酶。溶酶体的形成是由来自高尔基体的小泡与内涵体(endosome, E)融合而成的球形或不规则的结构。溶酶体内的pH值较低,主要依靠膜上的质子泵维持。溶酶体除降解蛋白质或脂类外,还参与膜损伤的修复过程。

溶酶体相关细胞器(lysosome-related organelle, LRO)是一类与溶酶体结构类似的,分布于特殊细胞中的细胞器,其形态各异。LRO的共同点是低pH值,含有溶酶体膜特异性标记蛋白(LAMP),如血小板中的致密体、黑色素细胞中的黑色素体、肺泡表面Ⅱ型上皮细胞的Lamellar小体、抗原提呈细胞的MHC池、神经元突触小体、中性粒细胞的嗜苯胺蓝体、自然杀伤细胞的溶解小体等。其形成涉及多个蛋白质复合体(BLOC-1,2,3)的参与。这些

复合体成分的缺陷，会导致 LRO 形成或功能障碍，引起多器官不同程度受累，如出血倾向、眼皮肤白化病、肺纤维化、免疫缺陷等，临床上被诊断为 Hermansky – Pudlak 综合征（HPS）、Chediak – Higashi 综合征等[10~12]，已建立针对 Hps 基因的专门数据库，参见http://falcon.roswellpark.org。

2.1.5.5 小泡转运（vesicle trafficking）

除线粒体和过氧化物体外，细胞内所有细胞器的生物合成所需的蛋白质和脂类都是从 ER 转运而来，其转运方式是通过小泡（vesicle）来完成。小泡实现货物运送一般经过出芽（budding）、锚定（docking）、融合（fusion）等过程，将来自供体的货物转运到靶向细胞器内，参见图 2 – 7。此过程涉及几十甚至几百种蛋白成分和一些有机或无机物（IP_3，Ca^{2+} 等）。目前认为，高等真核生物在进化过程中还具有其独特的、不同于低等真核生物或原核生物的转运途径，但其确切机制尚不明确[11,12]。

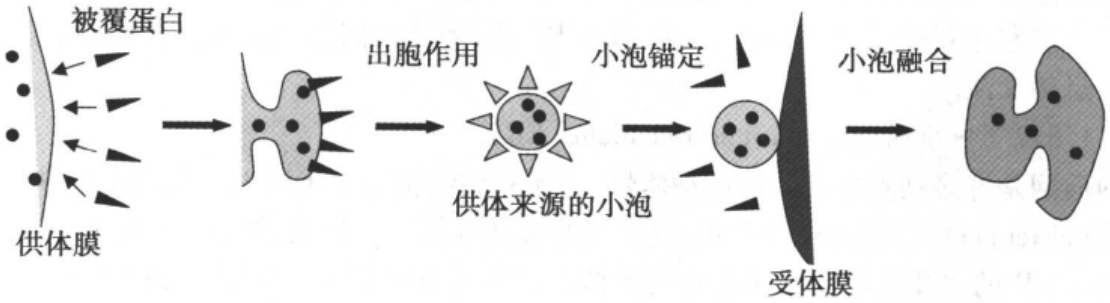


图 2 – 7 小泡形成中的出泡、锚定和融合过程示意图

图 2 – 8 总结了细胞内小泡转运的全过程[13]。小泡可分为两大类：一类无内涵素（clathrin）被覆；另一类有内涵素被覆。ER 和 Golgi 体之间的物质转运方式，目前尚有争议。两种无内涵素被覆，但有包被蛋白（coat protein，COP）的小泡参与其转运过程。COP Ⅱ主要介导 ER 到 Golgi 体的转运，COP Ⅰ则主要实现从 Golgi 体到 ER 的逆向转运。ER 中的管腔蛋白和膜蛋白一般分别带有 KDEL 和 KKXX（X 代表任何氨基酸）的特征结构，是货物运送的标签。溶酶体蛋白的标签是被甘露糖 – 6 – 磷酸（M6P）修饰，被溶酶体上的 M6P 受体所识别。

自 TGN 的外运渠道主要包括 3 类：一是通过小泡转运到质膜，形成质膜的组成成分，称为结构性转运途径（constitutive transport）；二是形成分泌型小泡，受激素等细胞内外因素调节进行出泡（exocytosis），将所携带的物质分泌到胞外，这种方式称为调节性转运途径（regulated transport）；第三种方式是形成有内涵素被覆的小泡，经过内涵体转运到溶酶体，这种途径称为内涵体转运途径（endosomal transport）。内涵体分早、晚期内涵体（EE 和 LE），LE 又被称做多泡体（multivesicular body，MVB）。内涵体转运途径中，涉及 3 种以上的内涵素被覆小泡，分别命名为 AP1，AP2 和 AP3 等。AP1 和 AP3 主要是从 TGN 到内涵体或溶酶体，经 AP3 运送的蛋白质通常具有双亮氨酸或富含酪氨酸的标签。AP2 小泡运送是从质膜经内涵体到溶酶体，也被称为胞吞作用（endocytosis）。经 AP2 途径的胞吞作用，首先在质膜形成被覆小凹（coated pit），与 AP2 小泡融合后，内涵素解离，剩余的小泡结构与内涵体融合，

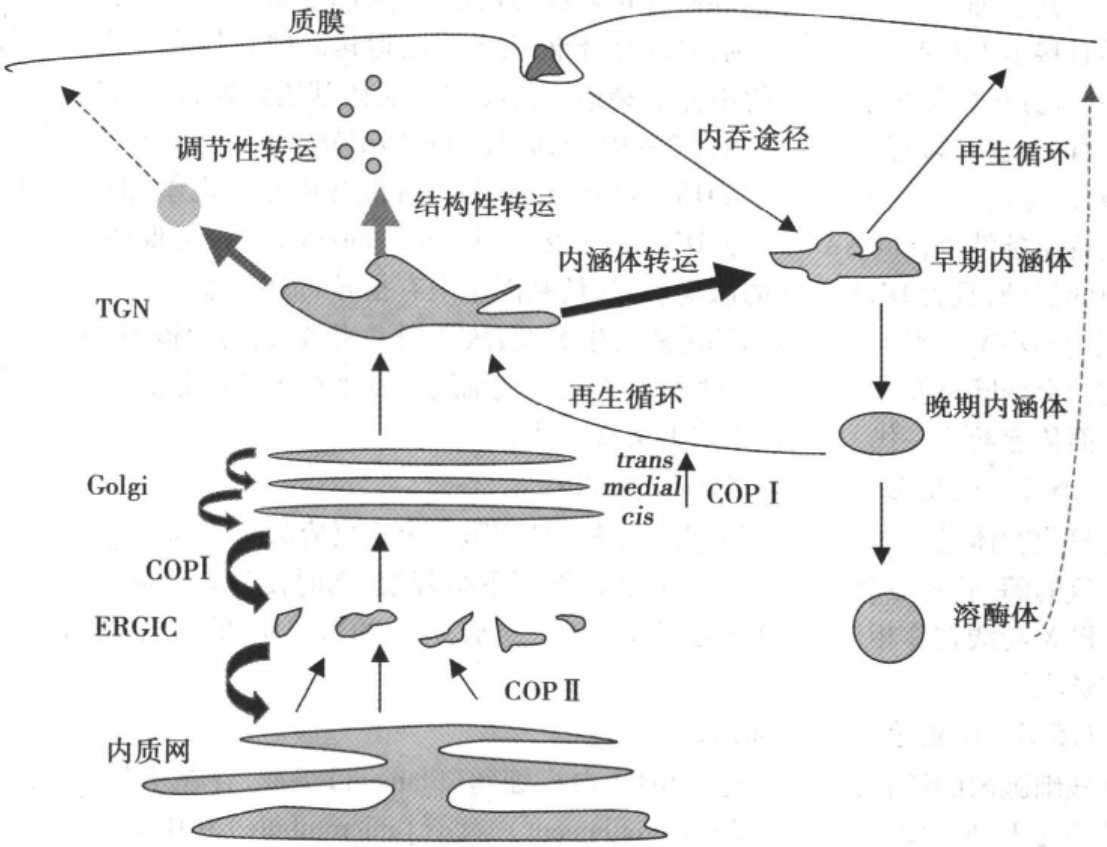


图2-8　小泡转运途径

后者参与溶酶体的形成。这种运送方式最经典的例子是 LDL 受体途径。另一种不依赖 AP2 小泡的胞吞作用是形成穴样结构(caveolae),直接将所包绕的货物运送到 Golgi 体或 ER,而不是内涵体或溶酶体,这也被称为胞饮作用(cell drinking)。

2.1.5.6　线粒体 (mitochondria)

线粒体普遍存在于需氧呼吸的真核细胞的细胞质中,是细胞内能量产生的场所,又称细胞动力工厂,是细胞内 ATP 合成和氧化磷酸化的场所,也参与细胞内 Ca^{2+} 的调节和细胞凋亡(apoptosis)的信号传导(细胞色素 C 的释放)。线粒体是细胞核外惟一含 DNA 的细胞器,其 DNA 组成类似原核生物。线粒体 DNA(mitochondrial DNA,mtDNA),又称线粒体基因组(mitochondrial genome),全序列在 1981 年由 Anderson 等测定,是一个含 16 569 bp 的双链闭环分子,外环为重链,内环为轻链,共编码 13 种蛋白质、22 种 tRNA 分子和 2 种 rRNA 分子。这 13 种蛋白质都是呼吸链酶复合体的亚单位。mtDNA 结构紧凑,不含内含子,惟一的非编码区为约 1 000 bp 的 D-环区,该区包括 mtDNA 重链复制的起始点和轻、重链的转录启动子以及 4 个高度保守序列。mtDNA 两条链的碱基组成差别较大,重链含 G 较多,轻链含 C 较多。重链编码 12 种蛋白质、12S rRNA、16S rRNA 和 14 种 tRNA;轻链仅编码一种蛋白质(ND6)和 8 种 tRNA。mtDNA 具有两个复制起始点,分别作为轻、重链的复制起点。转录启动子是从 D-环开始,重链转录按逆时针方向,轻链转录按顺时针方向。

尽管线粒体能进行自主复制、转录、翻译和分裂,但所编码的蛋白仅占线粒体蛋白质的

5%左右，其余都来自于核基因组编码、ER 合成的蛋白质，大部分脂类也来源于 ER。蛋白质和脂类直接由 ER 或 Golgi 体得到，其转运不依赖于小泡运送。线粒体膜分内膜和外膜两层，仅允许分子质量小于 5 ku 的小分子穿透。内膜向线粒体基质内伸，形成皱褶。每个细胞含数百个线粒体，每个线粒体内含 2～10 个拷贝的 mtDNA 分子，因此，每个细胞具有数千个 mtDNA 拷贝。同一细胞中的 mtDNA 可相互作用，出现混合或互补现象，但不发生重组。在某些特定条件下，mtDNA 可与核 DNA 发生交换或重组。mtDNA 因缺乏像核 DNA 那样的组蛋白的保护，且无 DNA 损伤的修复系统（核基因组中有指导 mtDNA 修复突变的基因），因此，每个 mtDNA 分子中每个碱基都可能发生突变，故突变率相当高，约为核 DNA 的 10 倍以上。这种突变既可发生在体细胞，也可发生在生殖细胞。发生在生殖细胞的 mtDNA 突变可随配子而传递给下一代，其遗传方式下文将述及。

2.1.5.7 过氧化物体（peroxisome，PEX）

过氧化物体是一种类似小泡的膜样结构，是细胞内氧化反应的主要场所，内含大量产生过氧化氢的酶，后者在触酶（catalase）的催化作用下被降解，同时使许多物质包括细胞毒物氧化。PEX 与线粒体相似，可自身分裂和繁殖。组成 PEX 的蛋白质和脂类直接在胞浆内合成而被送入。

2.1.5.8 细胞骨架（cytoskeleton）

真核细胞内的细胞骨架由一复杂的蛋白样细丝（filament）组成，存在于胞浆和胞核中。细胞骨架的 3 种主要形式为微丝（microfilament）、微管（microtubule）和中间细丝（intermediate filament，IF）。

微丝是由分子质量为 45 ku 的肌动蛋白（actin）组成的实体丝状结构，直径约 6 nm，占细胞总蛋白的 5%以上。主要发挥支撑和维系质膜的作用，同时参与细胞运动、维持细胞形态，被形容为“细胞的肌肉”。许多蛋白参与肌动蛋白由可溶性球状单体（即 g－actin）到丝状多聚体（即 f－actin）的聚合和解聚过程。

微管是直径约 25 nm 的中空管状结构，由微管蛋白（tubulin）组成。微管蛋白分 α－，β－和 γ－微管蛋白 3 种。微管的功能是维系胞浆和参与细胞内运动，并与许多动力蛋白（motor protein）协同作用，参与小泡和一些细胞器的移动。此外，微管在细胞分裂中也起重要作用，形成纺锤体；微管还参与纤毛上皮细胞的纤毛运动以及精子的鞭毛运动。

中间细丝（IF）是直径约 10 nm 的实体丝状结构，由一个或多个极其复杂的 IF 蛋白家族组成，存在于胞浆和胞核内，主要作用是加固胞浆和支撑胞核内膜。IF 具有组织特异性，不同组织的 IF 组成有所不同。IF 蛋白呈长条棒状结构，其中部结构域富含 α－螺旋结构。IF 蛋白的聚合由磷酸化介导，但高度磷酸化后可使 IF 解聚。

细胞骨架是细胞运动的物质基础，能量（如 ATP）的供应是细胞运动的动力来源。细胞运动的主要形式为细胞内运动、细胞变形运动（cell shaping）、细胞趋化运动（locomotion）、细胞泳动（swimming）、肌肉收缩运动等。

2.1.5.9 细胞核（nucleus）

细胞核是细胞内最明显的结构，被当做细胞的核心，含有细胞的核基因组。染色体是细胞核的基本物质，也是基因的载体。组成染色体的基本物质是 DNA 分子。一个 DNA 分子组成一条染色体。人类基因组由 24 种不同的 DNA 分子组成，总长度约为 30 亿个核苷酸。

染色体的基本结构单位为核小体(nucleosome)。每个核小体在由组蛋白组成的八聚体(H2A, H2B, H3, H4 各 2 分子)的基础上,外面盘绕 1.75 圈 DNA(约 140 bp),两个核小体之间以 60 bp 的 DNA 双螺旋与组蛋白 H1 形成的细丝相连,每 6 个核小体绕成一圈,形成空心螺线管。以此为基础,再进一步折叠和多级螺旋化形成在长度上被压缩了近万倍的中期染色单体。遗传信息的传递过程主要涉及细胞核的生物学功能。

细胞核由核膜包绕。核膜分内膜和外膜两层,核膜上有细小的核孔,通过被动或主动转运方式与细胞核外发生物质交换,转录形成的 mRNA 就是通过核孔进入核糖体,参与蛋白质的生物合成。

2.2　生物信息在细胞间的流向[2]

细胞分裂前,DNA 以半保留复制形式将 DNA 复制后,使遗传信息由亲代细胞传递到子代细胞。

2.2.1　细胞周期及其调控

细胞周期(cell cycle)是指两次细胞分裂之间的周期。在此期间,细胞必须复制其内容物并等量分配到两个子代细胞中,遗传信息也随 DNA 的复制和细胞分裂从亲代细胞流向子代细胞。细胞周期可分为 G_1, S, G_2, M 期 4 个阶段。S 期即合成期,DNA 的复制在该期完成。M 期即有丝分裂期,细胞在此阶段一分为二。位于 S 和 M 期之间的两个阶段分别称为 G_1 和 G_2 期,分别为 S 期和 M 期作准备。细胞周期的长短因细胞类型而异。一般地,分裂细胞 G_1 期为 6 ~ 12 h;S 期为 6 ~ 8 h;G_2 期为 3 ~ 4 h;M 期约为 1 h。细胞离开细胞周期后进入 G_0 期,也可从 G_0 期重新进入细胞周期。细胞周期过程参见图 2 - 9。

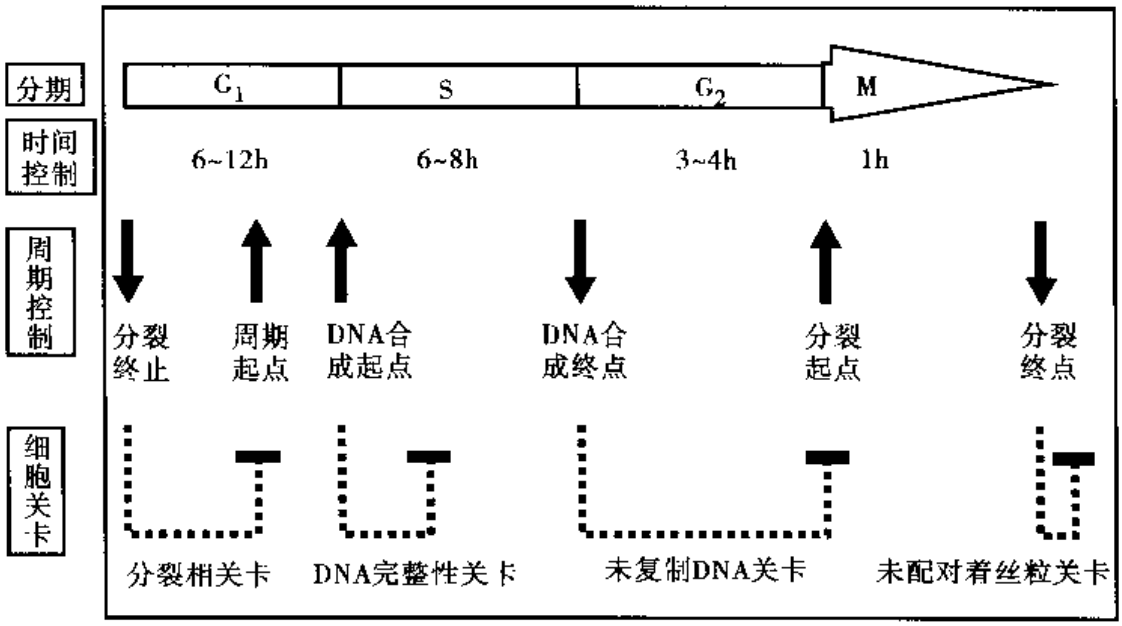


图 2 - 9　细胞周期及其调控

细胞周期调控的机制很复杂，涉及信号传导系统和多个细胞关卡（checkpoint）的调控。细胞受生长因子刺激后，通过信号传导系统产生一系列效应分子，引起基因的表达，促进细胞生长和分裂。细胞关卡主要控制着细胞周期的进程，两个主要的细胞关卡出现在 $G_1 \rightarrow S$ 和 $G_2 \rightarrow M$ 的交界处。当 DNA 受到损伤后，未修复的 DNA 难以通过 $G_1 \rightarrow S$ 的关卡进入 S 期；同样，当 DNA 复制未完成时，难以通过 $G_2 \rightarrow M$ 的关卡进入 M 期。细胞关卡的调控中最重要的参与者为一系列激酶，能使靶蛋白质发生磷酸化。这些激酶与周期蛋白（cyclin）结合后表现出酶活性，同时又受到一系列抑制物的控制。细胞正是在一系列严格调控机制的控制下完成一个细胞周期，同时保证遗传信息传递的精确性。当信号传导系统的基因发生突变时，常常导致细胞生长失控，从而导致肿瘤的发生，即所谓的癌基因；当控制细胞关卡的基因发生突变时，不能抑制细胞的生长，也可导致肿瘤的发生，即所谓抑癌基因。图 2－9 即为细胞周期及其调控示意图。

2.2.2 有丝分裂

细胞有丝分裂（mitosis）过程分为四期：前期（prophase）、中期（metaphase）、后期（anaphase）、末期（telophase）。G_1，S，G_2 期统称为间期（interphase），不分裂的细胞停留在该期。分裂细胞离开 G_2 期后进入有丝分裂的前期，染色质开始浓缩，形成姐妹染色体，核膜破裂。进入中期后核膜和核仁消失，染色体通过附着于着丝粒（centromere）的纺锤丝排列于细胞的赤道板。随着着丝粒的分离，细胞进入后期，姐妹染色体分离，分布于细胞两极，核膜开始围绕染色质形成，至细胞出现双核时，即进入末期，细胞沿赤道板形成缢坏，最终分裂成两个子代细胞，各自含有与亲代细胞完全一样的染色体。

2.2.3 减数分裂

减数分裂（meiosis）过程与有丝分裂有相似之处，又有明显的不同。减数分裂包括两次（Ⅰ和Ⅱ）连续的分裂过程，两次分裂之间不发生 DNA 的复制，产生 4 个子代细胞，每个细胞含亲代细胞染色体数目的一半，称单倍体（haploid）。减数分裂与有丝分裂一样，也被分成前期、中期、后期和末期 4 个阶段，但染色体的变化大不一样。

间期细胞进入前期Ⅰ后，经历较有丝分裂的前期复杂得多的过程，发生同源染色体的联会（synapsis），配对的姐妹染色体之间 DNA 双链断裂，遗传物质发生相互交换，再重新连接。中期Ⅰ的染色体分布于赤道板，与纺锤丝相连。进入后期Ⅰ，两套染色体各自移向两极，而不是靠纺锤丝牵拉向两极，这与有丝分裂的后期有本质不同。在末期Ⅰ被核膜包绕形成两个核，每个核所含染色体数目为母细胞的一半。末期Ⅰ的细胞分裂后不久即直接进行减数分裂前期Ⅱ，此期间未发生 DNA 复制。

减数分裂Ⅱ与有丝分裂过程相似，细胞又被一分为二，最终形成单倍体细胞。减数分裂Ⅰ和Ⅱ的示意图见图 2－10。值得注意的是，减数分裂过程中，来自父母双方的染色体发生了两次交换。一次是前期Ⅰ中联会后的交换；另一次是后期Ⅰ中，父源和母源的染色体随机组合后，移向两极，使子代细胞含有重组后父母双方的遗传信息。这使得生殖细胞的遗传信息并非拷贝单一亲代，从而保证了生物多样性。

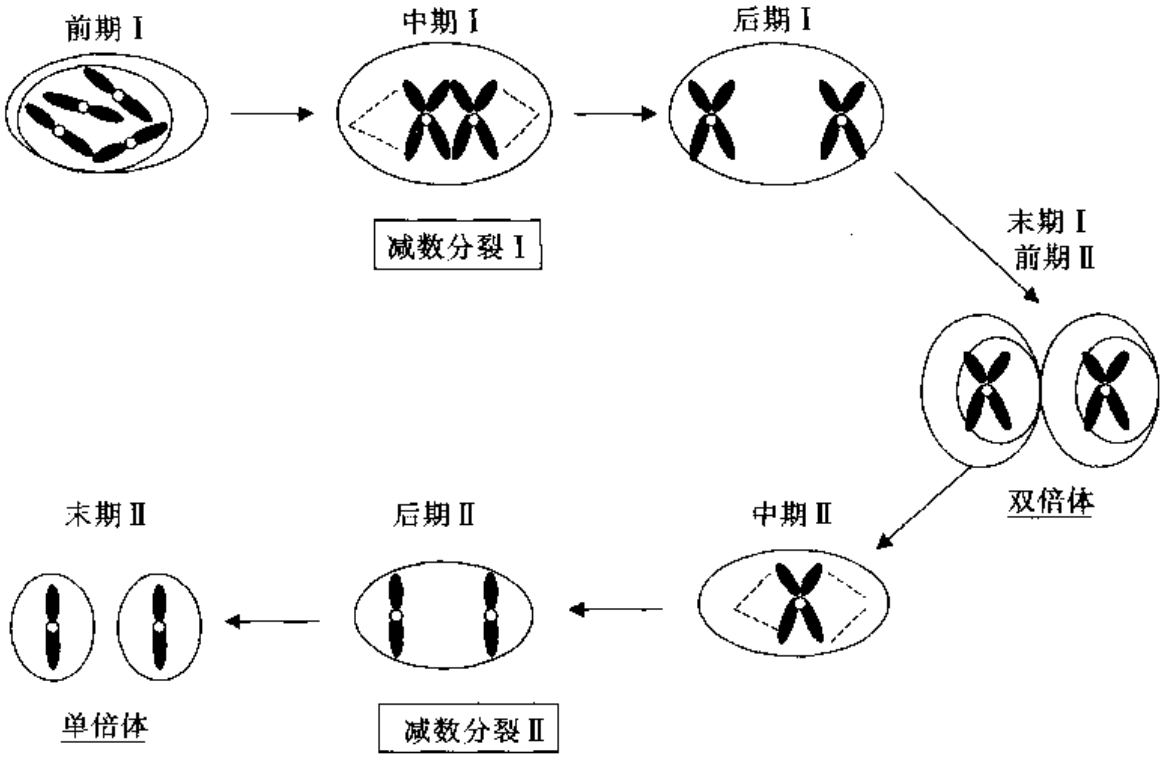


图 2-10　减数分裂Ⅰ和Ⅱ示意图

2.3　生物信息在世代间的传递

基因作为遗传的基本要素，遵循遗传规律代代相传。基因控制生物性状的遗传是非常复杂的。可以是基因产物的直接作用，也可以是通过基因产物作用于性状表达的中间环节而间接发挥作用，还可通过作用于基因表达的调控环节来控制性状的表达。对于多基因控制的性状，是由于不同基因产物的相互叠加作用来控制。

2.3.1　孟德尔遗传方式[2,3]

2.3.1.1　孟德尔基本规律

在生殖细胞形成过程中，等位基因彼此分离，分别进入不同的生殖细胞中，这一规律称为分离律，是由奥地利著名遗传学家孟德尔于 1865 年通过豌豆杂交实验所发现，又称孟德尔第一定律。100 多年来，这一规律被用来解释许多人类遗传病和性状的遗传规律。

孟德尔在总结一对相对性状遗传规律的基础上，进一步研究了两对以上相对性状的遗传，发现以下规律：两对或两对以上的等位基因位于非同源染色体的不同位点时，在生殖细胞形成过程中，非等位基因独立行动，可分可合，有均等机会组合到同一个生殖细胞中。这是由于在形成配子的减数分裂过程中，同源染色体要相互分离，非同源染色体随机组合进入不同的配子中。自由组合律又称孟德尔第二定律。

自由组合律主要针对非同源染色体上的非等位基因的遗传规律。但许多基因位于同一染色体上,这一现象称为基因连锁。1909 年美国遗传学家摩尔根及其学生在孟德尔定律基础上,利用果蝇进行的杂交实验,揭示了位于同源染色体上不同坐位的两对以上等位基因的遗传规律,即著名的连锁与互换规律。其基本内容是:生殖细胞形成过程中,位于同一染色体上的基因是连锁在一起的,作为一个单位进行传递,称为连锁律。在生殖细胞形成时,一对同源染色体上的不同对等位基因之间可以发生交换,称为交换律或互换律。

连锁(linkage)和互换(cross - over)是生物界的普遍现象,也是造成生物多样性的重要原因之一。一般而言,两对等位基因相距越远,发生重组的机会越大,即重组率(recombination rate)越高;反之,相距越近,重组率越低。因此,重组率可用来反映同一染色体上两个基因之间的相对距离。以基因重组率为 1% 时两个基因间的距离记做 1 厘摩(centimorgan,cM)。据此,可绘制出一条染色体上不同基因或多态标志物的遗传图(genetic map),这是人类基因组计划实施中的重要步骤。目前基于连锁与互换律,已广泛用于疾病基因的定位克隆和遗传病产前诊断中的连锁分析。

2.3.1.2 孟德尔遗传病分类和系谱分析

单基因遗传病中,根据决定该疾病的基因所在的染色体不同,其遗传方式主要分为以下几种。

(1)常染色体显性遗传(autosomal dominant inheritance),简称常显或 AD。

(2)常染色体隐性遗传(autosomal recessive inheritance),简称常隐或 AR。

(3)X 连锁显性遗传(X - linked dominant inheritance),简称 XD。

(4)X 连锁隐性遗传(X - linked recessive inheritance),简称 XR。

(5)Y 连锁遗传(Y - linked inheritance)。

(6)线粒体遗传(mitochondrial inheritance)。

临床上判断单基因病的遗传方式常用系谱分析法。系谱(pedigree)或称家图是指对某遗传病患者家族各成员的发病情况进行详细调查,再以特定的符号和格式绘制成反映家族各成员相互关系和发病情况的图解。系谱图的绘制方法常以该家系中首次确诊的患者又称先证者(proband)开始,追溯其直系和旁系各世代成员及该病患者在家族亲属中的分布情况。根据绘制的系谱图进行分析,又称系谱分析,以确定该家系是否患有遗传病及其可能的遗传方式。

系谱分析的基本程序是:先对某遗传病患者各家族成员的发病情况进行详细调查,再按一定方式将调查结果绘成系谱,然后根据孟德尔定律对各成员的表现型和基因型进行分析。通过系谱分析可以判断某种遗传病是单基因病还是多基因病,以及确定单基因病的遗传方式,探讨遗传异质性的存在。另外,系谱分析也是遗传风险分析、连锁分析和产前诊断中必不可少的工具。

2.3.1.3 常染色体显性遗传

一种性状或遗传病基因位于常染色体上,其性质是显性的,这种遗传方式就是常染色体显性遗传(AD)。所引起的疾病称为常染色体显性遗传病。

常染色体显性遗传具有以下特点[参见图 2 - 11(a)]:

(1)致病基因位于常染色体上,遗传与性别无关,男女发病机会均等。

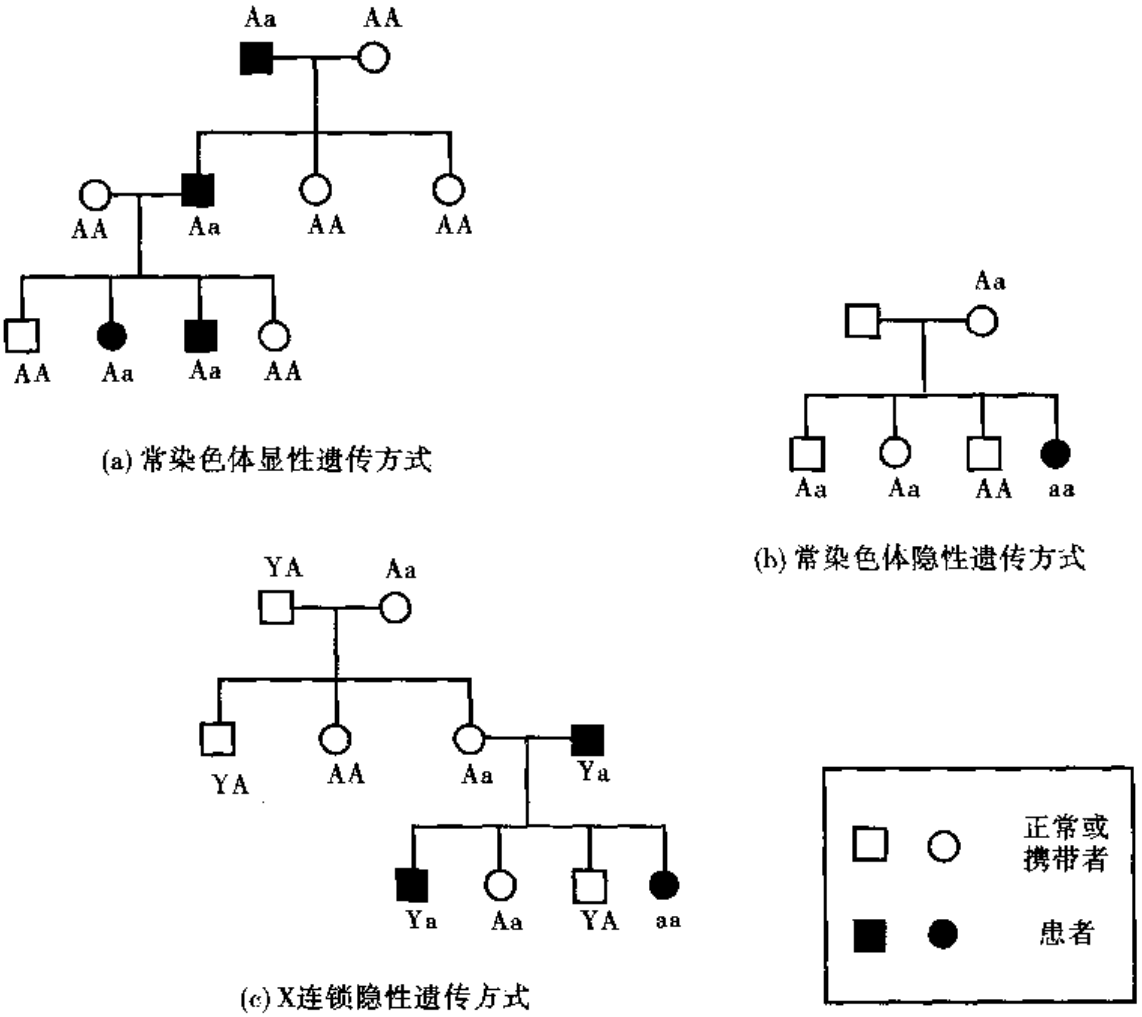


图 2-11 常见孟德尔遗传方式

(2)患者双亲之一为患者,绝大多数为杂合子,纯合子极少见。致病基因由患病亲代传来。双亲都未患病,子女一般不发病,除非新发生的基因突变,这种情况一般见于那些突变率高的病种,如多发性神经纤维瘤、软骨发育不全、成骨不全、家族性多发性结肠息肉、成人多囊肾等。

(3)患者同胞中约有 1/2 为患者,这在患者同胞数较多时明显。

(4)患者子代中约有 1/2 为患者,而且每次都有 1/2 生育一个患儿的可能。如果双亲同为患者(杂合子),则子代中,将有 3/4 为患者,仅 1/4 为正常。

(5)每代都可出现患者,出现连续传递现象。

2.3.1.4 常染色体隐性遗传

一种性状或遗传病基因位于常染色体上,其性质是隐性的,即在杂合状态下不能表现出相应症状,这种遗传方式就是常染色体隐性遗传(AR)。所引起的疾病称为常染色体隐性遗传病。这种具有隐性基因的杂合子(Aa)称为携带者(carrier)。

常染色体隐性遗传具有以下特点[参见图 2-11(b)]。

(1)患者双亲不发病而是肯定携带者。子女中出现患者、携带者、健康人的概率分别为 1/4、1/2 和 1/4。

(2)患者同胞中约有 1/4 患病,且男女患病机会均等。患者表现型正常的同胞,有 2/3 的可能性为携带者,这种携带者又被称为可能携带者。

(3)患者子女中,一般不发病,不出现连续传递现象。多为散发或隔代遗传。

(4)近亲结婚时,子女中发病风险增加。

2.3.1.5 X 连锁隐性遗传

致病基因位于 X 染色体上,随 X 染色体传递,称为 X 连锁遗传,分 X 连锁显性和隐性。X 连锁遗传不同于常染色体上基因的传递方式,对于 X 连锁基因而言,因男性只有一条 X 染色体,又称为半合子(hemizygote)。父亲 X 染色体上的基因只能传给女儿,男性 X 连锁致病基因只能从母亲获得,这称为交叉遗传。

当致病基因位于 X 染色体上,杂合时并不发病,称为 X 连锁隐性遗传(XR)。其特点如下[参见图 2-11(c)]。

(1)系谱中男性患者远多于女性患者,而且致病基因频率越小,男女发病率差异越大,甚至极少见到女性患者。

(2)双亲无病,即父亲正常,母亲为携带者时,儿子有 1/2 机会患病;女儿无患病风险,但有 1/2 为携带者。

(3)由于交叉遗传,患者兄弟、姨表兄弟、舅父、外甥有患病风险。

由于一些 XR 遗传病患者是致死性的,一般于婚育前死亡,这时很少见到连续传递现象。系谱中经常见到几代由女性携带者传递的方式。另外,一些致死性散发性 X 连锁隐性遗传病中,1/3 病例是由母亲卵子形成中新发生的基因突变引起。

2.3.1.6 线粒体遗传

线粒体基因组含有细胞核基因组之外的遗传信息,有其独特的遗传特点,表现如下。

(1)mtDNA 具有半自主性。这是指 mtDNA 能独立复制、转录和翻译。但维持线粒体结构和功能的多数蛋白质又受核 DNA 影响。

(2)线粒体基因组所用的遗传密码与核基因组的通用密码有所不同。主要的区别包括 UGA 编码色氨酸,而非终止密码。tRNA 的兼用性也较强,22 个 tRNA 可识别 48 个密码子。

(3)mtDNA 呈母系遗传(maternal inheritance)方式。即由母方将其 mtDNA 传递给所有子女,再由其女儿传给下一代。这是因为受精时,精子仅核 DNA 进入受精卵,其胞质不进入。

(4)mtDNA 具有异质性与均质性。mtDNA 在有丝分裂和减数分裂期间都要经过复制分离,加之 mtDNA 突变率高,使得体细胞和生殖细胞中可以同时具有突变型和野生型 mtDNA 分子,称为异质性(heteroplasmy)。异质性细胞经过有丝分裂和减数分裂后,随机分配到两个子代细胞,突变型和野生型 mtDNA 的比例发生改变,分别向纯合突变型或纯合野生型漂变,再经过多次分裂后,细胞达到完全纯合,称为均质性(homoplasmy)。异质性和复制分离现象表明,即使核基因组完全相同的个体,如一卵双生,也可具有不同的细胞质基因型,从而表型有所不同。mtDNA 突变一般以错义突变、tRNA 突变、插入/缺失突变、拷贝数目突变等 4 种类型的突变为常见。

(5)mtDNA 具有阈值效应。当突变的 mtDNA 达到一定比例时,才可出现一定的表型。表型严重程度具有组织特异性,即病情与该组织器官对氧化磷酸化的依赖程度有关,脑、骨

骼肌、心脏、肝脏等,对能量的依赖性渐次降低,一般受损程度也渐次减轻。

(6)mtDNA 的进化率极高。一般比核 DNA 高 10 ~ 20 倍。高进化率引起个体 mtDNA 序列在群体中的差别较大。这一方面可将 mtDNA 用于个体识别,同时也说明,线粒体病基因型很普遍,但表型的类型不多。

2.3.2 多基因遗传

2.3.2.1 多基因遗传概述

多基因遗传(polygenic inheritance)又称多因素遗传(multifactorial inheritance),是由多个基因的累加效应引起的遗传性状,一般与环境因素共同作用,所导致的疾病称多基因遗传病(简称多基因病)或多因素遗传病。因有遗传因素在内,故发病呈家族倾向,但不符合孟德尔遗传规律,即同胞中的患病率远比 1/2 或 1/4 低,只有 1% ~ 10%。

近年来,由于分子生物学的飞速发展以及基因检测技术的广泛应用,在诊断、风险率估计和治疗单基因病和染色体病方面已取得巨大进展。然而这些疾病只不过是人类遗传性疾病中的一小部分。大多数先天性畸形,如无脑儿、脊柱裂和其他神经管缺损以及大多数先天性心脏病,许多常见的成年性疾病,如癌症、高血压、冠心病、痛风、精神分裂症、抑郁症及糖尿病等,不是单纯由单基因突变或染色体异常所引起的疾病,这些疾病都是由多个基因和环境因素共同作用的结果,属于多因素遗传病。

2.3.2.2 数量性状

多因素遗传病为一种复杂的疾病,为了便于了解其发病特性及推算风险率,我们先了解一下多基因病的基础——数量性状(quantitative traits)。

数量性状的遗传基础也是基因,不过不是一对基因,而是两对以上的基因。各对基因呈共显性,每一对基因的作用是微小的,称微效基因。但是,若干对基因作用累加,可形成一个明显的效应,称为累积效应(additive effect)。近年来也发现一些主基因的作用所引起的多基因病。人类的身高、体重、血压是由多数基因的累加效应和环境因素共同作用的结果,由于这些性状可以用数字来测量,我们可以比较不同数量基因所引起的不同表型在群体中的分布情况。

以身高为例,首先假设一个人的身高是由在一个位点上的两个对子 A 和 a 所决定,A 基因使人个高,a 使人个矮,A 和 a 在人群中分布频数均等,且无显性与隐性之分。排列组合后群体中就可能会有 3 种基因型:AA,Aa,和 aa;从而产生 3 种表型:高个、中等和矮个。假设身高由两个位点上的 4 个对子所决定,第二个位点上的两个对子分别为 B(高),b(矮),如果 B 和 b 基因对身高的影响与 A 和 a 的作用相同,这时群体中会产生 aabb,aaBb,aaBB,Aabb,AaBb,AaBB,AAbb,AABb,AABB 9 种可能的基因型和 5 种表型。若身高由 3 个位点上的 6 个对子来决定,第三个位点上的基因是 C(高)和 c(矮),群体中则可能出现 27 种基因型,导致 7 种表型。依此类推,随着位点数目的增加,各表型频数分布图逐渐倾向于正态分布曲线。在此基础上,如再加上环境的影响,图形会变成更平滑的正态分布曲线。这就被称为连续性状(continuous trait)。事实上,任何由大量各自独立的基因联合作用所导致的可变性状在群体中都呈正态分布。

2.3.2.3 易患性、发病阈值和遗传度

在多基因遗传病中，遗传基础和环境因素共同决定一个个体患病的可能性大小，称之为易患性(liability)。易患性是多基因遗传中的特定概念。其中因遗传基础决定该个体的患病风险称为易感性(susceptibility)。易患性的变异也与上述数量性状一样，呈正态分布。在一个群体中，易患性高低不一，多数人处于中等水平，即接近平均值。当一个个体的易患性达到或超过一定的水平，就可能患病，这个正态分布曲线上发病的最低限度称做阈值(threshold)。这样，连续分布的易患性变异就被阈值区分出两个部分，即健康者和患者，使连续变异的数量性状在阈值部位发生了质的变化。在一定条件下，阈值代表患病所需的最少的有关基因的数量。

一个个体的易患性高低难以直接检测，但是，一个群体的易患性平均值可通过该群体的发病率进行估计。如果某病发病率高，说明群体中引起该病的易患性因素多，其易患性平均值距离发病阈值更近。相反，易患性平均值和发病阈值相距越远，表明平均值低而阈值高，群体发病率越低。因此可以从调查某群体中某病的发病率入手，计算出阈值与平均值的距离。该距离可用正态分布的标准差作单位。例如，当一种病的群体发病率为2.3%时，以阈值为零作估计，易患性平均值应该位于与阈值相距2个标准差的位置；如果一个群体中某病发病率是0.13%，阈值与平均值之间距离是3个标准差。

易患性的高低受遗传和环境因素双重影响，其中，遗传因素所产生的影响程度称为遗传度(heritability)或遗传率，通常用百分率(%)来表示。某病的遗传度高，表明遗传因素在该病易患性中所起作用大；遗传度低，则说明环境因素在起主要作用。多数遗传病的遗传度在60%～80%。遗传率的计算是根据一般群体和患者亲属易患性分布的对比而求得。一般通过查Falconer表得到相关数据再计算。因计算过程烦琐，此处不作赘述。

2.3.2.4 多基因遗传的特点

(1)多基因病有家族聚集倾向，所以患者亲属的发病率高于群体发病率，但在一个家庭中并没有明显的孟德尔遗传方式。

(2)多基因遗传病的发病风险与遗传度密切相关。根据群体发病率、遗传度和患者一级亲属发病率之间的关系，可以估计多基因病的发病风险率。当群体发病率为0.1%～1%时，遗传度如果是70%～80%，则患者一级亲属的发病率接近于群体发病率的平方根。当遗传度低于该值，患者一级亲属的发病率低于群体发病率的平方根。相反，遗传度高于此值，患者一级亲属的发病率高于群体发病率的平方根。例如，唇裂在中国人群体中的发病率为0.17%，其遗传度为76%，患者一级亲属的发病率为4%，近于1.7/1000的平方根(=4.1%)。如果遗传度为100%，患者一级亲属的发病率接近于9%；如果遗传度为50%，则患者一级亲属的发病率将低于2%。另一方面，在遗传度相同的情况下，群体发病率不同，发病风险率也不同。例如，在遗传度为50%的条件下，群体发病率为0.1%时，患者一级亲属的发病率为1%，即较群体发病率高10倍；群体发病率为1%时，患者一级亲属的发病率为5%，即较群体发病率高5倍；群体发病率为10%时，患者一级亲属的发病率为20%，即较群体发病率高2倍。

(3)亲缘关系的远近与发病率也有关系。患者一级亲属有相同的发病率，这与AR明显不同。二级亲属[叔，伯，舅，姑姨，侄(女)，外甥(女)]患病的危险性较一级亲属患病的危

险性明显下降,但其后远亲患病的危险性下降得较慢。例如,唇裂患者一级亲属发病率为4%,二级亲属为0.7%,三级亲属为0.3%。这一遗传特性有别于常染色体遗传病,在常染色体显性遗传,每一代的危险性都较前一代降低一半,危险性固定且容易推测。常染色体隐性遗传病中只有同胞患病,无其他亲属患病。

(4)家庭中若有一个以上的成员患病,再发风险率增高。例如在一个家庭中只有双亲之一患神经管缺陷,再发风险为4.5%;若双亲之一再加一个子女患病,再发风险增加到12%;若双亲之一再加两个子女患病,再发风险增加到20%。而单基因疾病不管有一个、两个或更多个病孩出生,再发风险对下一个孩子总是一成不变的。

(5)病情越严重,亲属的再发风险率越高。患病严重的个体表明其家庭具有更多的易感基因,所以再发风险就越大。比如单纯性唇裂患儿,其同胞再发风险为4.0;若患者患双侧唇裂和腭裂,其同胞再发风险增加到5.6%。这一点也与单基因遗传病不同,在单基因遗传病中,不论病情的轻重如何,一般都不影响其再发风险。

(6)当某个疾病在一种性别的发病率高于在另一种性别的发病率时,发病率低的性别其后代再发风险升高;相反,发病率高的性别,患者后代再发风险则较低。这是因为发病率低的性别,发病阈值高,一旦发病,则意味着其带有较多的致病基因。例如,先天性幽门狭窄的男性发病率高于女性5倍;女性患者的儿子中,发病率为20%;男性患者的儿子中,发病率为5%。

(7)双亲近亲婚配,子女再发风险率高。这是因为近亲婚配的双方带有更多相同的从共同祖先遗传来的致病基因。

2.4 生物信息流向紊乱与遗传病

2.4.1 基因突变的定义和分类[2,5,14]

突变(mutation)是指遗传物质(染色体或基因)因理化因素的影响或因DNA复制、细胞分裂过程中发生的可遗传的变异。这种变异通常改变了基因的DNA序列或染色体的结构或数目。染色体结构或数目的突变又称染色体畸变,导致染色体病。狭义的基因突变是指个别基因的结构所发生的变异。可表现为基因中某一位点碱基的改变,称点突变(point mutation)或基因大片段序列的改变(gross mutation)。

基因突变分类方法很多,根据突变的区域大小可分为点突变和大片段突变两大类。其中点突变又分为以下几种。

2.4.1.1 错义突变(missense mutation)

因碱基置换改变了密码子的编码性质,而编码出与野生型蛋白质不同氨基酸序列的蛋白质。这种突变通常发生在密码子的第一和第二个碱基。密码子第三个碱基的改变很少引起所编码氨基酸的不同。这种不引起编码氨基酸改变的基因突变又称为同义突变(same-sense mutation)或沉默突变(silent mutation)。

2.4.1.2 无义突变(nonsense mutation)

因碱基置换将密码子变成终止密码子,导致蛋白质的合成提前结束。这种突变所得到

的 mRNA 较不稳定，合成的蛋白质常因缺少 C－末端而无功能，引起较严重的症状。

2.4.1.3 移码突变(frameshift mutation)

由于一个或几个碱基(不是 3 个或 3 的倍数个)的插入或缺失，导致整个阅读框架(ORF)的改变，使得所编码的蛋白质在突变点以后的序列与野生型完全不同，甚至提早出现终止密码子。这类蛋白质通常失去正常生物学功能。

2.4.1.4 延长突变(elongation mutation)

因碱基置换或移码突变后，使原来的终止密码子变成某个氨基酸的密码子，而使肽链合成延续下去，直到下一个终止密码子出现，又称为终止密码突变。如中国人群常见的一种点突变型 α 地中海贫血——Hb CS (Constant Spring)。

2.4.1.5 拼接突变(splicing mutation)

因 mRNA 剪切加工时，内含子两端拼接识别序列(GT－AG)或外显子拼接增强子(ESE)的突变，导致正常的外显子拼接错误，常引起大片段的缺失或插入，而改变了所编码蛋白质的性质。另一种常见的拼接突变发生在内含子中间序列的点突变，产生一个可被识别的拼接信号，又称隐含拼接突变(cryptic splicing mutation)。

2.4.1.6 转录加工突变

因 mRNA 转录启动子、增强子、沉默子(silencer)、位点控制区(LCR)、多聚腺苷酸加尾信号等的突变，导致 mRNA 不能正常转录或加工，仅出现少量或完全缺乏具生物学功能的 mRNA。

图 2－12 为常见点突变方式的示意图[15]。

正常序列

ATG|AAC|CGT|CGC|CCG|TCA|CCG|TTA|TTG|CGT|

Met--Asp--Arg--Arg--Pro--Ser--Pro--Leu--Leu--Arg

沉默突变 ↓A→C

ATG|AAC|CGT|CGC|CCG|TCC|CCG|TTA|TTG|CGT|

Met--Asp--Arg--Arg--Pro--Ser--Pro--Leu--Leu--Arg

错义突变 ↓T→A

ATG|AAC|CGT|CGC|CCG|ACA|CCG|TTA|TTG|CGT|

Met--Asp--Arg--Arg--Pro--Thr--Pro--Leu--Leu--Arg

无义突变 ↓C→A

ATG|AAC|CGT|CGC|CCG|TAA|CCG|TTA|TTG|CGT|

Met--Asp--Arg--Arg--Pro--X

移码突变 ↓InsG

ATG|AAC|CGT|CGC|CCG|TCG|ACC|GTT|ATT|GCG|T

Met-Asp--Arg--Arg--Pro--Ser--Pro--Thr--Val--Iso--Ala--

图 2－12 常见点突变方式示意图

大片段基因突变包括缺失(deletion)、插入(insertion)和重排(rearrangement)等,导致基因结构的显著变化,所编码的蛋白质也因此失去正常生物学功能。引起大片段基因突变的原因有:拼接突变、转座子插入(transposonal insertion)、染色体错配与不等交换,以及因染色体易位导致的融合基因(如 BCR－ABL 基因)等。转座子(transposon)实际上是一种可移动的基因,可从染色体的一个位点移到另一个位点,或从一个复制子移到另一个复制子。已报道 30 多种基因因转座子的插入突变而导致相应的疾病。

近年来一类被称做动态突变(dynamic mutation)的现象引起临床遗传学家的关注。有些短串联重复序列(short tandem repeat, STR),尤其是三核苷酸重复,位于基因的编码或非编码序列中,它们的重复次数在一代代传递过程中会发生明显变化,从而导致某些遗传病的发生。如脆性 X 综合征、强直性肌营养不良、Huntington 氏病等。

最近报道一类新的遗传病产生方式,即反义 RNA 的转录可导致基因沉默(gene silencing)和基因调节区 CpG 岛甲基化(methylation),因此引起相关的疾病,如 α－地中海贫血[16]。

外遗传学(epigenetics)是近年来遗传学研究的热点之一。其性状是指体细胞性或生殖细胞性遗传物质的改变所导致的基因功能的改变无法由 DNA 一级结构的改变所能解释的遗传现象。外遗传学性状是可遗传的。这些现象包括基因组印记(genomic imprinting)、单亲二体性(UPD)、DNA 甲基化、基因表达的调节、X 选择性失活、转座现象等。

2.4.2　基因突变的效应[2,4]

根据基因突变对基因功能的影响,大致分为功能丧失(loss－of－function)、功能增加(gain－of－function)和功能不变 3 类。功能丧失通常因零突变(null mutation)引起,一般为隐性突变。功能增加常为显性突变,使蛋白质获得新的功能。基因突变的后果根据严重程度大致分为以下几类。

2.4.2.1　有利突变

突变给个体的生育或生存带来一定好处。

2.4.2.2　中性突变

突变对个体不产生可察觉的效应。

2.4.2.3　遗传多态现象(polymorphism)

突变对个体虽无危害,却是引起个体差异的重要原因。这种差异可体现在 DNA、mRNA、蛋白质、染色体等不同水平,分别称为 DNA 多态性、转录本多态性、蛋白质多态性和染色体多态性等。可以此作为基因定位、个人身份鉴定、药物反应性、疾病易感性、器官移植等的重要依据。因此,近年来有关 DNA 多态性的检测与致病基因突变位点的检测一样,受到临床上的广泛重视。

2.4.2.4　导致遗传病

突变通常导致蛋白质无功能,又称零突变(null mutation)。截至到 2003 年 10 月,已明确超过 1 500 个基因的突变可导致疾病效应。

2.4.2.5　致死突变

突变导致死胎、流产或死产。

2.4.3 遗传病的基本概念和分类[14]

临床医学中，遗传多态性和遗传病是涉及基因突变后果的两种最常见的现象。通常突变基因（mutant）与野生型基因（wild - type）位于染色体的同一位点，可视做等位基因（allele）。组成该个体的基因型（genotype），其所产生的突变效应，称为表现型（phenotype），可得到新的性状（trait）或疾病特征。突变基因与野生型基因同时存在于杂合体（heterozygote）中，突变基因的性状能表现出来，则称此为显性遗传（dominance）。突变基因的性状只有在纯合体（homozygote）中才表现，称之为隐性遗传（recessive）。

基因突变可发生于个体发育的任何阶段和体细胞（或生殖细胞）的任何分期，因此又区分为体细胞突变（somatic mutation）和生殖细胞突变（germline mutation）。体细胞突变仅在其子代细胞中传递，不遗传给后代。一般认为，多数肿瘤的发生属体细胞突变引起。只有生殖细胞的基因突变才能遗传给下一代，称为遗传病（inherited disease）。体细胞或生殖细胞基因突变引起的疾病统称为遗传性疾病（genetic disease）。

遗传性疾病大致分为以下几类：单基因病、多基因病、染色体病、体细胞遗传病、线粒体遗传病等。截至 2003 年 5 月 24 日，已在在线人类孟德尔遗传（OMIM，http://www.ncbi.nih.gov/omim）中收录的条目共 14 461 条，其中已明确的基因或表型位点为 10 731 种（常染色体占 10 104 种，X 连锁占 549 种，Y 连锁占 41 种，线粒体基因组占 37 种）。已知的遗传病种类为 2 233 种，其中已确定的疾病基因为 1 510 种。随着人类基因组计划的完成和疾病基因克隆技术的发展，以上这些数目将会不断增加。有关基因的定位除 OMIM 中可查找外，还可通过其他网站查找，如 NCBI 数据库中的 GeneMap99（http://www.ncbi.nih.gov/genemap），MORBID 图谱等。人类基因突变数据库（Human Gene Mutation Database，HGMD）则收录了常见基因突变类型（http://www.hgmd.org）。如果已诊断为某种遗传病，想要进一步了解是否可进行基因诊断，可查看http://www.genetests.org。

（李 巍 张 庆）

参 考 文 献

1 Kanehisa M, Bork P. Bioinformatics in the post - sequence era. Nat Genet, 2003, 33 (suppl): 305

2 李巍，何蕴韶. 遗传咨询. 郑州：郑州大学出版社. 2003.1 ~ 56

3 陆国辉，李巍，黄艳仪，等. 产前遗传病诊断. 广州：广东科学技术出版社. 2002.3 ~ 37

4 Lewin B. Genes VII. Oxford: Oxford University Press, 2000.1 ~ 346

5 Krawetz SA, Womble DD. Introduction to Bioinformatics: A Theoretical and Practical Approach. New Jersey: Humana Press, 2003.1 ~ 232

6 Champe PC, Harvey RA. Lippincott's Illustrated Reviews: Biochemistry. 2nd ed. Philadelphia: Lippincott - Raven Publishers, 1994.75 ~ 86

7 Pollard TD, Earnshaw WC. Cell Biology. Philadelphia: Saunders, 2002.381 ~ 671

8 Manning G, Whyte DB, Martinez R, *et al*. The protein kinase complement of the human genome. Science, 2002, 298: 1 912

9 Johnson GL, Lapadat R. Mitogen - activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinase. Science, 2002, 298: 1 911

10 Suzuki T, Li W, Zhang Q, *et al*. Hermansky - Pudlak syndrome is caused by mutations in HPS4, the human homolog of the mouse light - ear gene. Nat Genet, 2002, 30: 321

11 Zhang Q, Zhao B, Li W, *et al*. Ru2 and Ru encode mouse orthologs of the genes mutated in human Hermansky - Pudlak syndrome types 5 and 6. Nat Genet, 2003,33: 145

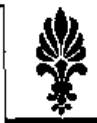
12 Li W, Zhang Q, Oiso N, *et al*. Hermansky - Pudlak syndrome - 7 (HPS - 7) is caused by mutant dysbindin, a new member of the BLOC - 1 complex. Nat Genet, 2003, 35: 84

13 Roche PA. Intracellular protein traffic in lymphocytes: "How do I get there from here?" Immunity, 1999,11: 391

14 Scriver CR, Beaudet AL, Sly WS, *et al*. The Metabolic & Molecular Basis of Inherited Disease (MMBID). 8th ed. New York: McGraw - Hill,2001.3 ~45

15 Guttmacher AE, Collins FS. Genomic medicine - A primer. N Engl J Med, 2002, 347: 1 512

16 Tufarelli C, Stanley JAS, Garrick D, *et al*. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet, 2003, 34: 157



第3章　人类基因组与基因组生物信息学

基因组(genome)是用来描述组成一种生物每个细胞内DNA的全部序列。不同生物具有其独特的基因组。人类基因组是目前已知基因组中最大、最复杂的基因组。对生物信息学工作者的基本要求是能掌握人类基因组的分析方法,这首先要求对人类基因组的组成特点有清晰的认识,也是基因组生物信息学(genome bioinformatics)的基本内容。基因组学(genomics)是指研究基因组的科学,其目的是阐明基因组的组成、基因的功能与相互作用,以及基因与疾病的关系等。随着人类基因组计划的完成,基因组学发展迅猛,目前已形成以下几个主要的分支[1]。

(1)解剖基因组学(anatomic genomics)　主要研究基因组的解剖结构,是人类基因组计划(HGP)的基本内容,包括序列图、全长cDNA及基因组多样性(SNP图)等。

(2)结构基因组学(structural genomics)　进入后基因组时代(post-genomic era),结构基因组学主要研究核酸或蛋白质的结构、定位、功能及其相互作用;与蛋白组学(proteomics)内容密切相关。本章仅对人类蛋白组计划作一般介绍,其他有关结构基因组学内容详见第6章。

(3)功能基因组学(functional genomics)　主要研究基因的表达、调控、功能及基因间的相互作用。第7章"微阵列生物信息学"是功能基因组学的重要内容。

(4)比较基因组学(comparative genomics)　研究不同模式生物(model organism)基因组与人类基因组的关系,以更好认识人类生理、病理过程和行为方式等,见图3-1。其中还包括对不同基因组的比较研究,认识基因或蛋白的进化规律,以便了解蛋白家族的形成和相关功能研究。这也是进化遗传学的主要内容,详见第8章。

(5)药物基因组学(pharmacogenomics)　研究药物对全基因组表达的影响以达到药物筛选的目的。常涉及药物遗传学(pharmacognetics)的内容,即有关药物个体反应性或药物遗传性状、药物代谢的遗传控制等。药物基因组学还涉及研究药物分子设计和分子靶向的化学基因组学(chemical genomics)的内容,将在本书第9章作相关介绍。

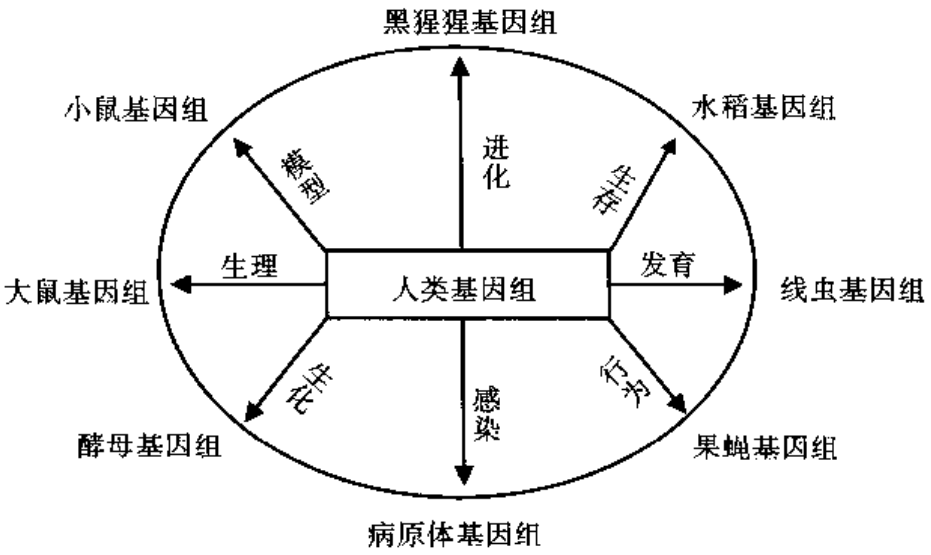


图 3－1　比较基因组学的研究对象及相互关系

(6)医学基因组学(medical genomics)　或称基因组医学(genomic medicine),是研究人类及其相关的基因组在医学中的应用,主要探讨基因与疾病的关系、基因型与表现型的关系、遗传因素与环境因素的相互作用,以及基因组资源在医学中的应用等。包括通常所说的疾病基因组学、病原体基因组学、应用基因组学等。基因组医学与其他临床学科一样,有自己独特的研究"器官",即基因组,因而已成为临床医学的一个新的分支学科。

近年来,基因组学与生物信息学的分支发展迅速,这两大学科各分支间存在着相互联系,参见图 3－2。本章将重点叙述结构基因组学、比较基因组学和医学基因组学的相关内容。涉及生物信息学的层面,则重点介绍三大公共人类基因组数据库的应用。这三大公共基因组数据库为:①美国的 NCBI (http://www.ncbi.nlm.nih.org);②美国的 UCSC (http://www.genome.ucsc.edu);③欧洲的 Ensembl (http://www.ensembl.org)。

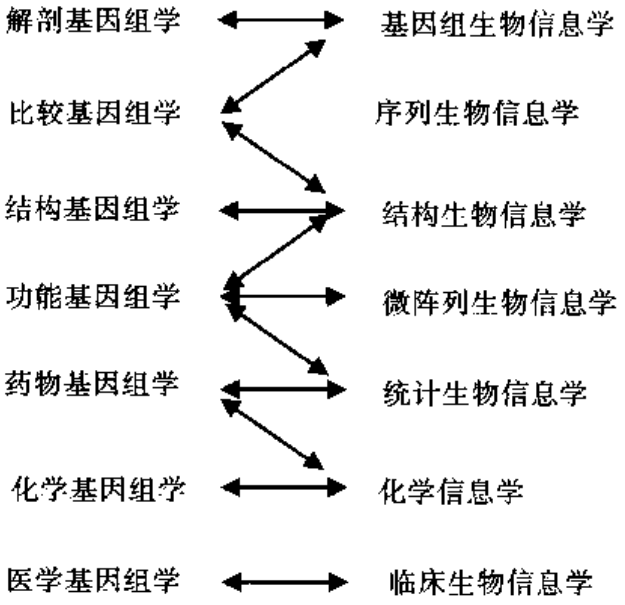


图 3－2　基因组学与生物信息学分支学科的相互联系

3.1 人类基因组计划:基因组的解剖结构

3.1.1 人类基因组计划的实施过程[2,3]

为了更好地认识人类基因组,有必要了解人类基因组计划(human genome project, HGP)的实施过程。HGP 的提出可追溯到 20 世纪 80 年代中期。1985 年当美国加州大学 Santa Cruz 分校(UCSC)的 Robert Sinsheimer 提出应该讨论对人类基因组测序的可能性时,许多人认为这种观点近乎疯狂或者不成熟。到了 1986 年,美国能源部(DOE)的 Charles DeLisi 决定对基因组定位和测序提供资助,同时诺贝尔奖得主、著名分子肿瘤学家 Renato Dulbecco 博士在《科学》杂志撰文论述其可行性和必要性。1988 年,由美国科学院下属的研究院特别小组向美国国会建议开始起动 HGP,计划用 15 年时间,每年投入约 2 亿美元,共计投入 30 亿美元测定人类基因组约 30 亿对碱基序列,其耗资及规模与曼哈顿原子弹计划及阿波罗登月计划相当。1988 年底,另一诺贝尔奖得主、DNA 双螺旋结构发现者 Watson 博士领导的一个美国国家卫生研究院(NIH)和 DOE 联合小组对 HGP 的实施起了有力的推动作用。1990 年美国国会通过了 HGP 的预算案,并于同年 10 月 1 日正式启动。90 年代初期 HGP 的其他几位关键人物是 NIH 国家人类基因组研究所(NHGRI)负责人 Francis Collins、Wellcome Trust 的 Michael Morgan 和 DOE 的 Aristides Patrinos。随着技术发展的需要和更多的国际组织与科学家的参与,一个行之有效、分工合理的国际人类基因组计划协作组(HGP consortium)成为 HGP 实施的核心领导。中国人类基因组研究中心成立于 1998 年 8 月,于 2000 年 4 月顺利完成 3 号染色体短臂上 30Mb 区域的工作草图,2001 年 9 月完成序列图,即所谓"1%项目",成为继美、英、法、日、德后加入到这一国际协作组 20 个测序中心的一员。

HGP 的实施大致经历了遗传作图→物理作图→定位测序→工作草图→序列图等关键阶段。对各阶段进行了严密的分工和目标设定,最终于 2003 年 4 月,正好是 DNA 双螺旋结构发现 50 周年之际,宣告 HGP 的提前完成(工作草图已于 2000 年 6 月 26 日宣告完成)。这一项目的实施过程堪称所有研究项目的典范。表 3-1 为 HGP 的项目计划和完成时间表[2]。表 3-2 为 HGP 大事年表[3]。

表 3－1　人类基因组计划的项目计划和完成时间表

研究项目	研究目标	完成指标	完成时间
遗传作图	2～5cM 分辨图(600～1 500 分子标记)	1 cM 分辨图(3 000 分子标记)	1994 年 9 月
物理作图	3 万个序列标签位(STS)	5.2 万个 STS	1998 年 10 月
DNA 序列图	测定 95% 的基因序列,准确性达 99.99%	完成 98% 以上的基因序列测定,准确性达 99.99%	2003 年 4 月
测序速度和成本	年测序量为 500 Mb,每个碱基成本 < $0.25	年测序量 > 1 400 Mb,每个碱基成本 < $0.09	2002 年 11 月
基因鉴定	全长 cDNA	1.5 万个全长 cDNA	
模式生物	完成大肠杆菌、面包酵母、线虫、果蝇的基因组序列测定	除完成所计划的 4 种模式生物基因组测序外,还完成了小鼠、大鼠等基因组的测序	2003 年 4 月
功能分析	建立基因组规模的功能分析技术	高通量寡核苷酸合成技术	1994 年
		DNA 芯片	1996 年
		cDNA 文库	1996 年
		面包酵母全基因组敲除	1999 年
		大规模酵母双杂交定位	2002 年

表 3－2　人类基因组计划大事年表

年 份	主 要 事 件
1984 年	首次公开讨论人类基因组测序项目
1986 年	首个依靠定位克隆的基因——DMD 基因被发现 DNA 自动测序仪问世
1987 年	第一代人类遗传图问世(RFLP)
1988 年	美国国立研究院提交"人类基因组定位和测序"项目计划书 人类基因组组织(HUGO)成立 酵母人工染色体(YAC)文库建立
1990 年	人类基因组计划正式启动 NIH－DOE 联合小组设立伦理、法律和社会问题(ELSI)研究项目
1991 年	美国基因组研究中心成立
1992 年	建立第二代人类遗传图(STS) NIH－DOE 联合小组提出基因组数据共享指南
1993 年	在《科学》杂志发表新的 HGP 五年计划(1993～1997 年) 英国 Sanger 测序中心成立(后改名为 Wellcome Trust Sanger 中心)
1994 年	完成遗传作图

续表 3-2

年份	主要事件
1995 年	完成物理作图
	通过基因非歧视法案
1996 年	首张人类遗传图问世
	人类基因组测序初步研究开始
	完成面包酵母基因组测序
	完成小鼠遗传作图
	公布基因组数据快速公开的百慕大原则
1997 年	DOE 成立联合基因组研究所(JGI)
	美国国家基因组研究所(NHGRI)成立
	完成大肠杆菌基因组测序
	法国国家基因组测序中心(Genoscope)成立
1998 年	绘制 3 万个遗传位点的遗传图
	在《科学》杂志发表新的 HGP 五年计划(1998～2003 年)
	日本 RIKEN 基因组科学中心成立
	完成线虫基因组测序
	启动单核苷酸多态位点(SNP)作图计划
	中国人类基因组中心成立
1999 年	大规模基因组测序开始
	完成首条人类染色体(第 21 号染色体)的测序
2000 年	完成人类基因组工作草图(6 月)
	完成果蝇基因组测序
2001 年	发表人类基因组初稿
	完成 10 万个人类全长 cDNA 的测序
2002 年	完成和发表小鼠基因组初稿
	完成和发表水稻基因组初稿
	完成大鼠基因组初稿
2003 年	宣告人类基因组计划完成(4 月)

值得一提的是,HGP 的提前完成,技术上主要归功于大规模测序技术的建立(如 ABI3700 测序工作站)、细菌人工染色体(BAC)文库的使用(主要为 Roswell Park 肿瘤研究所构建的 RPCI-11 文库),以及生物信息学序列组装技术等。与国际人类基因组计划协作组采用的染色体分区分段测序策略不同,美国 Celera 公司由 Craig Ventor 领导的人类基因组测序小组自 1998 年 5 月开始采用全基因组"鸟枪法"(shotgun)随机测序再计算机集成排序的策略,于 2000 年 6 月同期宣告工作草图的完成,这同样得益于大规模测序技术的发展和

生物信息学技术的发展与应用，所不同的是应用了鸟枪法随机测序技术[5]。

3.1.2 后基因组时代的任务和目标[3,6]

后基因组时代的任务主要分以下四大领域。

3.1.2.1 基因组与生物学(genome to biology)

阐明基因组的结构与功能。

(1)补平缺口　由于存在一些难测序区域，人类基因组尚存在约2%的缺口(gap)，预计尚需花10年左右的时间才能补平这些缺口，需要更新现有的测序技术才能完成。

(2)遗传多样性研究　主要是SNP研究、单体型作图(haplotype mapping)与比较基因组研究。收集不同人种样本，根据SNP等分子标记构建单体型，用于疾病基因的定位和连锁分析，进一步了解基因变异的种系发生学、人类学和生物学功能。同时通过比较基因组学研究，了解基因的演化机制和生理功能。

(3)逐条完成人类24条染色体的组装与分析　截至2003年5月，已完成22,21,20,14,7号染色体的全序列分析[7]。最近NHGRI已启动DNA元件百科全书(ENCODE)计划，以期透彻了解基因组中结构与功能的关系。

(4)了解基因与蛋白的功能　在代谢组和细胞组水平进一步了解基因与蛋白质的功能。

3.1.2.2 基因组与健康(genome to health)

将基因组信息转化，为人类健康服务。

(1)个体化医学与药物基因组学　通过对个体基因组分析，进一步了解基因变异导致个体对疾病的易感性或抵抗性，对药物的敏感性或耐药性。随着测序技术的发展，可望用较低成本对各个个体的全基因组进行测序，或建立基于SNP的DNA芯片对各个个体进行筛查，特别是针对那些编码区的沉默突变，利用统计所谓信息学方法分析SNP与易感基因或药物性状的相关性。

(2)基因预测医学和预防医学　通过基因检测进行症状前诊断，尽早干预或采取相应的预防措施，将大大提高生命质量和人类寿命。蛋白组学研究也为疫苗研究提供更多的选择。

(3)疾病基因组学与化学基因组学　人类基因组计划的完成已大大加速疾病基因的定位克隆，也从单基因病过渡到多基因病基因的克隆，这将为疾病的基因诊断、产前诊断、基因治疗提供更简便、快捷、有效的方案。了解致病基因的发病机制、代谢途径以及结构特点，作为药物的新靶点，以求寻找更有效、更安全的药物，是化学基因组学的重要内容。

3.1.2.3 基因组与社会(genome to society)

趋利避害地利用基因组资源，使之服务于全社会。

(1)新政策保障基因资源的合理利用　妥善利用基因资源将是各国政府面临的重要课题。某些遗传病基因的群体筛查需要相应的政府决策才能更好地为群体的健康或人口质量服务，但又必须符合伦理道德规范。

(2)基因组与种族、民族的关系　这是了解民族多样性的生物学基础，但需要有基因非歧视(genetic nondiscrimination)法案保障就业和医疗保险的公平权利。

(3)基因与行为的关系研究　将为犯罪学和社会学提供更大的挑战。遗传与环境相互作用的研究为政府制定环境保护措施提供理论依据。

(4)反生物恐怖　基因组资源也是一把双刃剑,服务于人类健康和人类社会的同时,其滥用也可造成对人类的危害。随着人类基因组的解密和疾病基因组的研究,使得恐怖分子可以利用人类或病原体基因组作为目标进行生物恐怖活动,生物安全将是全球面临的严峻挑战。

3.1.2.4　基因组与生命(genome to life)

利用环境微生物世界造福于人类。

(1)基因组分析　解析环境中关键微生物的基因组结构,系统分析各基因的功能,在蛋白组学水平了解其分子作用机制。进一步了解生物圈中各模式微生物的调节网络、各微生物之间的相互联系及其与环境的相互作用规律。

(2)生物信息学分析　整合所有测序的微生物基因组信息,建立相应数据库,并加以模拟分析和利用,解决相关的环境污染、气候变化与能源短缺问题,造福于人类。

3.1.3　人类基因组的结构特点[4,5,8]

人类基因组包括两个相对独立而又相互关联的基因组:核基因组和线粒体基因组。人类基因组全长约30亿碱基对,分布于24条不同的染色体或线粒体上。线粒体基因组全长约16 kb,参见2.1.5.6节有关线粒体的介绍。因线粒体基因组相对较小,通常人类基因组是针对其核基因组而言。人类基因组中,具有编码功能的基因及其基因相关序列约占25%,其中所有外显子序列的总和约占基因组的1%;其余75%序列为基因间序列(intergenic DNA),又称基因外序列(extragenic sequence)或无用DNA (junk DNA),其功能尚不清楚,可能参与基因表达的调节,通常含较多的重复序列。基因在染色体上的分布并不均匀,在染色体的某些区域基因含量丰富,称为基因"绿洲",被基因稀少的"沙漠"分隔。有关基因的结构参见2.1.1节。

3.1.3.1　基因数目

关于人类基因组所含的基因数目一直存在争议,最近普遍认为约3万个基因[4,5]。以人类7号染色体为例[7],其全长为157 953 789 bp,其中含1 455个编码基因(包括已知的、新的、部分性的和预测的蛋白)和213个非编码基因。依此类推,人类基因组含有约2.9万个蛋白质编码基因和3 700个非编码基因。由于选择性转录本的存在,实际存在的蛋白质种类大大超过3万个。例如人类7,14,22号染色体的已知基因中,存在选择性拼接的基因分别占55%,54%,59%。基因的大小各异,小的不超过1 kb,大的可超过100 kb,基因的平均长度约为50 kb。

3.1.3.2　基因重复序列

人类基因组中,约42%的基因序列为重复序列(repetitive DNA),呈串联状或分散状分布于整个基因组,是人群基因组多样性研究、进化遗传分析、染色体稳定性研究、基因多态性分析和用做DNA身份鉴定的重要区域。重复DNA可分为两大类:原位扩展重复序列和转座成分(transposable element, TE)。最常见的重复方式是串联重复(tandem repeat),又称简单重复序列(simple sequence repeat, SSR)。

另一种分类方法是将重复 DNA 分为卫星 DNA 和分散重复序列两大类。卫星 DNA 又称高度重复序列，根据重复片段大小分为微卫星 DNA、小卫星 DNA 和 α－卫星 DNA。通常将重复单位小于 10 bp 的称为微卫星 DNA，重复单位大于 10 bp 小于 70 bp 的称为小卫星 DNA，α－卫星 DNA 重复片段通常在 70 bp 以上。微卫星和小卫星 DNA 是基因组作图和染色体连锁分析的重要标志；α－卫星 DNA 主要位于染色体着丝粒附近，被认为与染色体联会有关，目前被广泛用做染色体荧光原位杂交(FISH)的探针。

分散重复序列常分为短分散重复序列(SINE)和长分散重复序列(LINE)。前者重复单位一般为 90～500 bp，如常见的 Alu 重复序列，多位于基因丰富区，占整个基因组的 10% 左右；后者重复单位可长达 7 kb，如常见的 L1 重复序列。这些分散重复序列可发挥转座子(transposon)的作用，容易引起 DNA 重组或基因突变的发生。转座子为基因组中可移动的序列，可插入到基因组不同部位。转座子通常分为两大类：一类为直接利用 DNA 序列编码蛋白质(转座酶)并在基因组中繁殖；另一类类似逆转录病毒的繁殖方式，利用 RNA 逆转录为 DNA 后，再插入到基因组中。这类转座子在基因组的插入部位两端通常具有相同的重复序列，与转座子的长末端重复序列(LTR)相连。

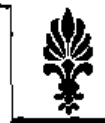
目前已建立专门的重复 DNA 数据库，称之为 Repbase Update(RU)（http://www.girinst.org），已收集约 2 400 种不同真核生物的重复序列。基因重复序列的鉴定和分析方法，主要依据 Smith－Waterman 运算法则，常用两个网站进行分析：http://www.girinst.org/Censor_Server.html 和http://repeatmasker.genome.washington.edu。基因组中有关转座成分的鉴定流程为：用 Repeat Masker 封闭已知重复序列→得到相应的互补序列→分别查找正向和互补序列的同源序列→进行多序列比对(multiple alignment)和通用序列(consensus sequence)分析→确定靶向部位和分类。

3.1.3.3 基因组复制

基因组复制(genome duplication)是人类基因组的普遍现象，是基因家族和同系基因(paralog)产生的重要原因。基因组复制主要包括两类：节段性复制(segmental duplication)和逆转座现象(retrotransposition)。节段性复制是基因的一大段或整个基因序列的重复。人类基因组中，出现节段性复制的区域超过 1 000 个，涉及的基因 3 500 多个。有趣的是，这种节段性复制所产生的同系基因，其中之一常为致病基因。另一种复制方式属于逆转录病毒的转座插入方式。这种复制产生的同系基因通常不含内含子，而是出现多个拷贝现象，随机分布，并且相似性极高，常导致基因诊断中较难克服的假阳性。

3.1.4 人类基因组的多样性[4,9]

平均而言，两个不同个体的基因组存在 99.9% 的相同性，其差异大致相当于每 1 kb 长度 DNA 中存在 1 个碱基的不同。如果这种碱基差异在人群中所占比例超过 1%，通常称之为多态性(polymorphism)。但应注意，在某些孤立或特殊人群中出现的高发遗传病的基因突变，有时突变等位基因频率也超过 1%，此时不应当做多态性对待，而应视做遗传病基因。人类基因组中，最常见的 DNA 变异是单核苷酸多态性(single nucleotide polymorphism，SNP)。据统计，人类基因组中，平均每 1 250 个碱基中就有一个 SNP。这些 SNP 主要出现在基因间序列，也可出现在内含子或外显子非编码区，编码区 SNP 相对较少，多不引起氨基



酸的改变。这种位于基因编码区或调节区的SNP以及RNA基因中的SNP的生物学意义相对重要,是药物基因组学的研究重点。已经启动的SNP单体型作图,是基因连锁分析、基因定位、个体识别、进化遗传学分析的重要工具。

3.2 人类蛋白组计划[9,10]

2001年10月,一个旨在研究人类蛋白组的民间性国际组织——人类蛋白组组织(human proteome organization, HUPO)正式成立(http://www.hupo.org)。该组织由北美、欧洲和亚太三大分支机构共同组成。中国属于亚太HUPO的成员之一。HUPO的各项计划在分步实施之中。目前主要启动三大项目:血浆蛋白组计划(plasma proteome project, PPP)、人类肝脏蛋白组计划(human liver proteome project, HLPP)以及蛋白质标准化数据库(protein standards initiative, PSI)。HUPO也在考虑实施其他组织或器官的蛋白组计划,如人类脑蛋白组计划(human brain proteome project, HBPP)等。蛋白组计划的实施需要借鉴HGP的成功经验,需要组织全世界一流的蛋白组学研究队伍、良好的组织管理与信息交流,以及不断发展的适应高通量和自动化需求的新技术和数据处理方法。

3.2.1 血浆蛋白组计划

该计划的总体目标是综合分析正常人体血浆和血清的蛋白组分,并比较血清与血浆样本的差别;比较全球不同人群间或同一国家内不同人群间血浆蛋白的差异;鉴定疾病状态下血浆或血清蛋白标记物。目前已有14个不同国家43个实验室参与此计划。PPP的实施需要建立标准化的操作方法和技术平台;主要是参考血浆的使用和保存、蛋白质分离分析方法以及如何去除含量高的常见蛋白质以便于微量蛋白质的分析。

3.2.2 肝脏蛋白组计划

HLPP的任务是:综合分析正常或疾病状态下肝脏蛋白质的组成;了解蛋白质的定位(又称定位组,localizome);综合分析蛋白质的相互作用和代谢网络图(又称相互作用组,interactome);建立PPP与HLPP的相互联系,找出相应的生物标记物;解析肝脏蛋白组与肝脏转录组和人类基因组的内在联系。目前HLPP主要是信息收集阶段,尚未进入实施。

3.2.3 蛋白组标准化数据库

与上述蛋白组计划相配合,需要建立大型数据库,以供结果分析和数据比对。这要求建立统一的数据输入方式,分门别类地管理,同时便于自由交换数据。目前欧洲生物信息学研究所(EBI)已经启动这一蛋白组标准化数据库项目。首先是针对蛋白组学两大关键领域的数据,即质谱和蛋白质相互作用,建立相应的数据管理和分析方法。

3.3 比较基因组学[9]

比较基因组分析是通过比较人类基因组与其他模式生物基因组的异同，以便更好地了解人类基因组的结构和便于蛋白组学研究。模式生物也是生理生化分析、进化分析、疾病研究、行为与发育研究等的重要工具，参见图 3-1。这些模式生物的基因组相对较小，分析方便。人类基因组计划完成的同时，小鼠、大鼠、河豚、果蝇、斑纹鱼、水稻、面包酵母、按蚊、疟原虫、线虫、多种细菌和病毒（如 HIV 和 SARS）等基因组的测序也相继完成，各基因组序列参见http://www.ncbi.nih.gov/PMGifs/Genomes/allorg.html。

通过对常见病原体基因组的分析，有利于了解病原体与宿主的关系、病原体的毒力、宿主对病原体感染的抵抗力和易感性等。同时通过对病原体基因组的变异分析，找出其流行规律。如 2003 年春季在我国广州、北京、香港以及新加坡、加拿大多伦多等地流行的 SARS 病毒，通过对基因组的分析，了解其进化关系和变异性，找出其他动物宿主，以便掌握其传播方式。更主要的是通过比较基因组分析，了解与其他冠状病毒的差异和毒力产生的原因，并通过对该病毒蛋白组分析，为疫苗的研究打下基础。攻克 SARS 的难关需要病毒学家、流行病学家、基因组学家、蛋白组学家、免疫学家和生物信息学家的通力合作。

小鼠的基因组成与人类很相近，因其易于繁殖，是遗传学、基因组学和蛋白组学研究的重要材料。目前已建立许多疾病的小鼠模型。这些模型大大推动了疾病基因的克隆、疾病分子机制探讨、基因治疗研究和药物研究等。从小鼠研究所得结果应用到人的过程，离不开比较基因组学的研究。大鼠的生理与人类较接近，也是常用的模式生物。

3.4 基因组医学及相关的社会、伦理、法律问题[1,11~14]

HGP 的完成是生命科学中划时代的里程碑，因此诞生了临床医学的一门新的分支科学——基因组医学。HGP 的完成也是 21 世纪的重要生产力，将对医学、生物制药、工农业生产带来重大变革，随之带来许多相关的社会学、伦理学、法律学的问题。

3.4.1 预测医学的发展趋势

人类疾病的发生归根结底是由于基因在作祟。可以认为，除理化损伤外，人类所有的疾病都是“基因病”，包括结核病、艾滋病等外源因素（感染）所致疾病也与人体的易感基因或多态位点有关。表 3-3 所列为常见疾病的相关基因。这些相关基因的检测作为症状前诊断依据是今后预测医学的重要内容。此外，通过检测药物反应的相关基因或多态位点，以指导临床用药，也是预测医学的内容，详见 9.2.5 节。

表 3-3 常见疾病的相关基因

疾病名称	相关基因	疾病名称	相关基因
老年性痴呆	APOF4	骨质疏松病	VDR
非胰岛素依赖型糖尿病	$PPAR_{\gamma}$/SUR1	哮喘病	ADRB2
胰岛素依赖型糖尿病	HLA-DQ_{α}	精神分裂症	NRG1/5-HT2A
低 HDL 性冠心病	ABCA1	男性不育症	线粒体 DNA
动脉粥样硬化	CETP	血色素沉着症	HLA-G
心血管病	FAT/CD36	乳腺癌/卵巢癌	BRCA1/BRCA2
原发性高血压	ACE/AGT	结肠癌	MSH1/MSH2/MSH3
左心室肥厚	PPARα	AIDS	CCR5

3.4.2 应用基因组学的发展

随着人类基因组计划的完成，基因资源在临床医学中的应用已日益广泛，主要体现在以下几个方面。

3.4.2.1 疾病的基因诊断与产前诊断

据估计，人类基因组约含 3 万个基因，截至 2003 年 5 月，已明确 1 510 个基因的突变可导致相应的疾病，主要为单基因病。理论上，这些被阐明的基因突变，都可建立相应的基因诊断方法，并用于产前诊断。

3.4.2.2 药物个体反应性评价

个体化医学(individualized medicine)将针对每个个体疾病表型的特点、药物反应性等进行合适的诊断和治疗，而不像传统医学中那样，对于任何一种诊断方法或治疗药物都是通用的。药物个体反应性主要基于个体的单核苷酸多态性(SNP)。研制一种新药时，须同时研究药物敏感型或抵抗型相关的 SNP，通过检测这些 SNP，指导合理用药。参见 9.2.5 节。

3.4.2.3 DNA 身份鉴定与亲子鉴定

由于基因组的多样性，DNA 的多态位点即成为个体识别的标签。依照遗传学规律，子女的身份标记必定来自于双亲，据此可用于身份鉴定或亲子鉴定。目前最常用做个人身份的 DNA 标签是一种 DNA 短串联重复序列(STR)，美国 ABI 和 Promega 公司相继推出一套 16 个 STR 位点的检测系统，用于个人识别和亲子鉴定，其亲权关系排除率可达 99.98% 以上。这套系统也用于公安刑侦中的重要法医物证鉴定、罪犯数据库建立、异基因骨髓移植嵌合度鉴定、医学差错鉴定等。目前也在试用 SNP 作为个体识别的依据，如对 9·11 事件中世贸中心残留物身份鉴定，特别是对可能已降解的 DNA 标本，仍有可能分析那些扩增片段大小在 100 bp 以内的 SNP。

3.4.2.4 植入前遗传学诊断与胎儿选择

植入前遗传学诊断(preimplantation genetic diagnosis, PGD)是在体外授精(in-vitro fertilization, IVF)基础上，从卵裂期胚胎发育到 6~10 个细胞阶段，采集 1~2 个细胞对遗传病

进行诊断的一种方法。PGD 成功的关键是胚胎活检和单细胞 PCR 或 FISH 技术的建立。胚胎活检技术是在显微操作下，将胚胎进行预处理使胚胎外周的透明区变薄，再用纤细的玻璃吸管刺入胚胎内，采集 1～2 个细胞，经单细胞 PCR 扩增后，可对某些单基因病进行诊断，或通过 FISH 技术对染色体病进行诊断；将确诊为正常的胚胎移植到子宫内。国内外已成功地开展多种遗传病的植入前遗传学诊断，如 α 地中海贫血、囊性纤维化、HbS 病、马凡综合征、唐氏综合征、DMD、范可尼贫血等。目前此法因技术要求高且体外授精所需费用昂贵，尚未得到普及应用，预计将有更多的疾病通过 PGD 技术进行有效的胎儿选择。

3.4.2.5 基因治疗

随着人类疾病分子生物学研究的深入及 DNA 重组与基因转移技术的发展，导致了基因治疗(gene therapy)方法的诞生，并从理论上使某些遗传病特别是单基因病得以根治。基因治疗是指运用重组 DNA 技术，修复患者体内有缺陷的基因而恢复该基因正常的功能，以达到治疗该遗传病的目的。包括基因修正、基因抑制、基因添加等方案。基因修正是定点导入外源正常基因，代替有缺陷的基因。基因抑制是应用反义核苷酸、干扰 RNA(RNAi)或核酶(ribozyme)抑制体内过度表达的基因，如某些癌基因。基因添加是非定点导入外源基因，而没有真正去除或修复缺陷基因，但导入体内后，可表达具有正常功能的蛋白而发挥治疗作用，相当于酶的替代疗法。药物基因治疗是指应用某些药物来诱导或抑制某些基因的表达。基因治疗的临床广泛应用取决于转基因技术的安全性和有效性。

3.4.2.6 基因制药与基因疫苗

利用基因重组技术，将某些具有治疗或预防价值的基因克隆到一定的载体中，通过体外或体内表达，得到相应的药物或疫苗。基因重组技术自 20 世纪 70 年代诞生以来，目前已在新药及疫苗的研究与开发中发挥极其重要的作用。

3.4.2.7 转基因动物

广义的转基因动物主要指利用基因转移技术培育的新的动物品种的方法，包括以下几个方面：一是转基因动物模型。即将外源基因导入(knock－in)动物基因组使其成为特定的动物模型。例如，利用乙肝病毒基因导入小鼠，可复制出乙肝的转基因小鼠模型，用于医学研究。二是基因敲除(knock－out)动物。即将动物中某种特定的基因通过基因操作技术，使其失活，或者通过基因诱变(mutagenesis)使某种基因失活而产生的动物。主要用于研究基因的功能与疾病表型或复制动物模型。三是动物乳腺生物反应器。是指利用原核显微注射(pro－nucleous microinjection)技术，将外源基因定向导入动物中，使其在乳腺中表达，产物可在乳汁中分离得到。这种转基因动物可成为某些药物的生产“工厂”。例如，将人白蛋白或促红细胞生成素(EPO)基因导入山羊或奶牛中，使其在乳腺中分泌相应的白蛋白或 EPO。四是克隆动物。通过对体外授精和胚胎发育过程的认识，将受精卵的胞核剔除后，移入体细胞的胞核，并使之发育为个体的过程，称为核移植，或生殖性克隆(reproductive cloning)，可在哺乳动物中完成无性生殖的过程。这就是 1997 年英国罗斯林研究所创造的轰动全球的克隆羊“多莉”的诞生过程。目前国内外科学家相继克隆了其他物种，如牛、鼠、鱼、猴、马等。这种克隆技术虽然给生物学研究、动物品种的保存和繁殖、疾病治疗等带来巨大的潜力，但与此同时，其应用也引起广泛的伦理和社会方面的争议。由于这种技术应用于人体，使得“克隆人”在技术上成为可能，由此引发“生殖性克隆”和“治疗性克隆”的广泛伦

理学争论和相关法规的制定。

3.4.3 遗传咨询与遗传伦理学的新范畴

人类基因组计划的完成丰富了遗传咨询和遗传伦理学的内容。遗传咨询(genetic couseling)是由咨询医师(counselor)和咨询者(counselee)(遗传病患者本人或其家属)就某种遗传病在家庭中的发生情况、再发风险、诊断和防治上所面临的问题,进行一系列的交谈和讨论,使患者或其家属对该遗传病有全面的了解,选择最适当的决策的全过程。一般可分为婚前咨询、生育咨询、一般咨询等。遗传咨询过程中,必须遵守的三大原则是:非指令性原则、尊重原则和保密原则。通过对人类基因组和遗传病病理的认识,近年来遗传咨询中非经典孟德尔遗传现象[或称外遗传学(epigenetics)现象]已引起广泛关注,同时新技术的不断涌现和应用,使遗传咨询师必须在咨询过程中作出合理的解释和利用。

遗传伦理学(genetic ethics)是伦理学的一个分支,主要阐明有关遗传检测技术应用和遗传资源利用的伦理、道德和社会方面的问题。遗传伦理学以个人权利、家庭利益和社会利益作为讨论的中心。由于遗传性疾病的终身性、难治性、可遗传性等特点,可能由此引发一系列的心理和伦理问题。随着人类基因组计划的完成,这种基于人类基因资源的遗传服务(genetic service)已得到越来越广泛的应用,相关的遗传伦理学问题的争论也将越来越激烈。例如有关胎儿选择、克隆人、干细胞治疗、症状前诊断、基因治疗、基因歧视、基因武器等。目前许多国家已出台相关的法律法规加以规范。早在1990年,人类基因组计划(HGP)已专门设立了"伦理、法律和社会问题"(Ethical, Legal, Social Implications, ELSI)专题研究组(http://www.nhgri.nih.gov/ELSI),召集各路专家如医学遗传学家、伦理学家、历史学家、理论家、法学家、政策分析员、社会学家等,主要研讨有关HGP相关的伦理、法律和社会问题,制定相关政策和法规。ELSI首要解决以下4个方面的问题:遗传信息的隐私权和公平利用;新的遗传工程技术应用于临床的安全性和有效性;遗传研究相关问题;公众及职业教育。随着人类基因组资源的广泛应用,遗传伦理学也将面临更多严峻的课题和挑战。

3.5 基因组图、全基因组扫描和定位克隆

3.5.1 基因组图[14]

基因组图(genomic map)是利用一定的界标(landmark)将基因组DNA制图,便于基因的定位分析和了解基因间的连锁关系。最常见的3类基因组图为:细胞遗传图(cytogenetic map)、遗传图(genetic map)和物理图(physical map)。目前在三大公共基因组数据库中可以见到这3种不同图谱的整合图。

3.5.1.1 细胞遗传图

细胞遗传图是通过特殊的染色方法将各染色体分成不同的区带,按照国际人类细胞遗传学命名委员会(ISCN)的命名原则,从染色体着丝粒开始沿染色体的臂向远端开始标记区和带。命名某个区带时按以下顺序:①染色体号;②臂的符号(p或q);③区号;④位于该区

内的带号;⑤带号后以小数点分隔出亚带号。例如 10p24.2 表示 10 号染色体短臂 2 区 4 带 2 亚带。这便是细胞遗传学中的核型分析(karyotyping)和染色体病诊断的基础。对于特定的 DNA 片段或基因,可以通过荧光原位杂交(FISH)技术,将其定位在染色体的特定区带位置。

3.5.1.2 遗传图

遗传图又称连锁图(linkage map),根据减数分裂过程中同源染色体的连锁或交换现象,在后代中可观察到两个比邻的分子标记可以出现连锁或分离现象。通过计算其重组的概率,得出相邻分子标记的距离。通常以重组率(以 θ 表示)为 1% 时(即平均每 100 次减数分裂中观察了 1 次交换现象),遗传图距为 1 厘摩(centimorgan, cM),大致相当于 1Mb 的物理图距。遗传图中所用的遗传标记有多种。最早使用的是蛋白多态标记,如血型、血清蛋白等,其后发展到 DNA 标记。第一代 DNA 标记为基于限制性内切酶谱建立的限制性片段长度多态性(RFLP)标记。串联重复 DNA 的发现,进一步过渡到可变数目串联重复片段多态性(VNTR)标记,即小卫星 DNA,属第二代 DNA 标记。短串联重复片段多态性(STR)标记或微卫星 DNA 的使用,使遗传作图更精细,这是第三代 DNA 标记。随着基因组计划的完成,现已开始使用单核苷酸多态性(SNP)标记,即第四代 DNA 标记绘制遗传图。SNP 的查找参见 SNP 数据库(http://www.ncbi.nlm.nih.gov/SNP)。遗传图是物理作图和定位克隆的基础,也有利于发现重组热点。

3.5.1.3 物理图

物理图是用来描述界标间的物理距离,类似高速公路上的里程标记牌。最早用做物理图的界标是 RFLP。PCR 发明后,主要使用序列标签位(sequence - tagged site, STS)作为界标,STS 的检索可查找http://www.ncbi.nlm.nih.gov/dbSTS。STS 除 STR 外,主要来自于表达序列标签(expressed sequence tag, EST)和 BAC 克隆末端序列的定位结果。人类基因组计划实施过程中,曾使用两种常用的 STS 作图法,即酵母人工染色体克隆(YAC)STS 作图和辐射杂交作图(radiation hybrid mapping, RH),参见http://www.ncbi.nlm.nih.gov/genemap99。随着人类基因组计划的完成,DNA 序列图是最精确的物理图,其中根据 cDNA 文库得到的 EST 序列也常被用做物理作图,又称转录图。EST 数据库参见http://www.ncbi.nlm.nih.gov/dbEST。一种与 EST 类似的分析方法被称为基因表达序贯分析(serial analysis of gene expression, SAGE),利用 cDNA 中 9 ~ 10 个碱基作为标签,比较不同组织特定基因表达量的差别,参见http://www.ncbi.nlm.nih.gov/SAGE。

3.5.2 全基因组扫描[9]

全基因组扫描(whole genome scan, WGS)是疾病或性状相关基因分析的重要方法,一般用于具有家族史的家系样品分析。也可用于缺乏家系资料而只有针对群体的病例对照(case - control)资料。目前 WGS 已广泛用于单基因病、多基因病或药物遗传性状的分析。WGS 的实质是对涵盖全基因组的分子标记进行的连锁分析,找出其中具有最大相关性的分子标记,并假设所寻找的相关基因与该分子标记紧密连锁,即具有连锁不平衡性(linkage disequilibrium, LD)。相关性一般以似然性比值的常用对数,即 Lods 值表示,其计算方法参见 8.5.2 节。因此 WGS 分析的关键在于样品资料和分子标记的选择以及 Lods 值的计算。

3.5.2.1 样品资料

关于样品资料，如果是具有家族史的家系资料，尽量采集患者一级亲属（双亲、同胞、子女）的血样，并有完整的病例记录。如果是基于群体的资料，常采用回顾性研究，在与患者相同群体中设置对照人群。一般要求样品资料足够大而且均匀。对于具有遗传异质性的疾病，应注意鉴别诊断并加以排除，其存在会大大降低连锁或关联分析的效力。如果存在人群混杂或分层（stratification）现象，应注意所选病例或对照的分层均衡性，如种族、籍贯、生活习性和环境等。对于群体研究中的对照人群应进行 Hardy－Weinberg 平衡检验，证实该对照人群处于平衡状态。

3.5.2.2 分子标记

目前常用的分子标记为微卫星 DNA。NCBI 数据库中已建立相应的微卫星 DNA 分子标记数据库（UniSTS），记录该分子标记的序列和定位以及相应的 PCR 引物。美国 Research Genetics 和 ABI 公司等已遴选出几套适用于 WGS 的分子标记，用做不同染色体覆盖密度的扫描，并且可选择荧光标记的 WGS 试剂，便于在全自动测序仪上使用，达到高通量和快速准确筛选，同时还有配套 Lods 值计算与作图软件。

微卫星 DNA 分子标记（即 STR）的缺点是其在染色体上的分布不均匀，多位于基因间序列，而且在基因重复区也较常见，可出现非特异扩增产物而对结果分析产生偏差。此外微卫星 DNA 的突变率也较高。相反，SNP 多态位点分布相对均匀，有些位于基因编码区，平均每 1 250 个碱基中含有 1 个 SNP，便于精细作图；而且其突变率低，在世代传递中较微卫星 DNA 稳定；只要选择非同系基因的 SNP，其序列特异性高，不易出现非特异扩增产物。因此可以预见的是，不久的将来，SNP 将会代替微卫星 DNA，成为新一代 WGS 的基本工具。

据估计，人类基因组中约含 2.4 百万个 SNP。国际人类基因组协作组已于 1998 年启动 SNP 的测定计划，目前已基本完成，并收录于 NCBI 的 dbSNP 数据库中。如何从中选择用于 WGS 的 SNP 是目前讨论的重要议题，需要综合考虑以下几点。

（1）SNP 的连锁不平衡性　LD 的表示方法是用 D' 值表示，界于数值 0（无连锁不平衡）和 1（完全连锁不平衡）之间。WGS 中所指的 LD 通常是指相邻两个 SNP 呈完全连锁不平衡，即其中一个 SNP 可 100% 地预测另一个 SNP。这些邻近 SNP 紧密连锁形成小区（block），遗传学上称之为单体型（haplotype）小区，被视做一个遗传单位。确定这些单体型的方法首先是通过统计学的最大似然性（maximum likelihood）运算法则来粗略估计，最终需要依靠基因型分析来确认。

SNP 单体型一般为 5 ~ 150 kb，同一单体型大小在不同人种中有一定差异，因此国际 SNP 协作组（TSC）于 2001 年启动 SNP 单体型作图计划，通过收集不同人种标本绘制出单体型图，理论上每一小区中的任何 SNP 都可代表该区域。重组率越高，LD 区间越小，即 SNP 单体型小区越小。LD 区间大小在染色体上并不均匀。影响 LD 区间大小或重组率的因素有：性别（男性重组率约为女性的 2 倍）、染色体臂（染色体短臂重组率较高）、与着丝粒的距离（离着丝粒越远，重组率越高），以及 GC 含量、重组热点、自然选择等。一般认为，WGS 中有用的 SNP 其 D' 值必须大于 0.5，即所选的两个 SNP 相距应小于 50 kb。若以建立在 SNP 单体型基础上的 SNP 为选用标准，平均相距 30 kb，则用做 WGS 的 SNP 共需 10 万个。

（2）SNP 的位置　约 99% 的 SNP 位于基因间或内含子，通常与疾病没有直接关系，最多

是连锁关系，而且这些部位多出现重复序列，可影响基因型分析。因此 SNP 的选择最好是基因密集区，但也不能完全遗漏基因荒漠区。一般认为那些位于基因编码区且引起氨基酸改变的 SNP 其临床意义最大，其次是那些位于基因上游调控区的 SNP，另外位于拼接点附近的 SNP 被认为可导致拼接改变，产生不同的转录本。通常将位于编码区的 SNP 称做 cSNP，每个基因平均含 2～4 个这种 cSNP。这样，人类基因组 3 万个基因中，cSNP 多达 12 万个，若检测其中的 10%，则需要检测 1.2 万个 SNP。估计约 40% 的 cSNP 可导致氨基酸改变，即人类基因组中有将近 5 万个 cSNP 可导致氨基酸的改变。还有人建议将全基因组扫描改为全基因扫描，即只针对 3 万个基因中的 SNP 进行筛查。

(3)SNP 的频率　约 75% 的 cSNP 其基因频率小于 15%，平均约 7%。这些频率低的 SNP，其多态信息含量也低，常不被选做 WGS 的 SNP。这些低频率 SNP 在筛查人群数量少时，特别易被忽略，但其中仍可能具有重要的生物学意义。

综上所述，根据 SNP 的分布特点和临床意义，用于 WGS 的 SNP 最少需要 3 万个，即相当于每个基因 1 个 SNP，覆盖全基因组的平均距离为 100 kb。如果分析成本可降低到 1 元/SNP，分析 100 个样品也需要 3 百万元。检测成本是目前阻碍 SNP－WGS 方法应用的原因，有赖于高通量仪器和低成本检测技术的发展。样品混合(pooling)是降低成本的方法之一，但不适合单体型分析。

3.5.2.3　数据分析

WGS 数据分析的基本方法是进行连锁分析，即计算各分子标记的 Lods 值并绘制 Lods 图。Lods 值的计算可用 SAS 软件(参见第 10 章)或专用 WGS 基因型 Lods 值分析软件(如 Allegro)[15]，也可用目前遗传流行病学中最常用的软件 GENEHUNTER 进行分析。如果样品具有完整的家系资料，通常利用传递不平衡检验(transmission disequilibrium test，TDT)来计算高风险单体型，这需要患者双亲的样品基因型分析。某些特殊情况下，如中老年性疾病(老年性痴呆、癌症等)患者双亲不在世而得不到样品，此时可通过受累同胞对(affected sibling－pair，ASP)分析单体型的风险率，参见 8.5.3 节。

对于结果的判断，一般认为当 Lods 值大于 3.0 时，具有临床或生物学意义。如果样本的干扰因素或相互作用因素多，即使 Lods 值较低，也不能忽视其生物学意义。解析 WGS 分析结果时，首先考虑 Lods 值最大的位点，但最好具有重现性。WGS 完成后，进一步对该区域进行精细作图，一方面验证 WGS 的结果，同时对候选基因精确定位，便于筛选。在未出现 SNP 的 WGS 试剂前，作者建议先用微卫星 DNA 标记进行 WGS，再检索 dbSNP 数据库，利用 SNP 作精细作图，缩小基因关键区。

3.5.3　定位克隆

目前致病基因的阐明主要通过两种途径：一是从表型入手，利用定位克隆(positional cloning)方法，确定致病基因(参见http://genome.nhgri.nih.gov/clone)。二是对于新发现的基因，通过人工突变(mutagenesis)，观察表型的变化，以明确基因与疾病的关系。候选基因研究(candidate gene studies)也常用于疾病或药物反应型基因的克隆，主要根据性状表现或药物作用靶向来确定候选基因，进行筛选。这种方法成功率较低，如能结合定位方法，则成功率大大提高，称之为定位候选克隆方法(positional candidate cloning)，是目前疾病基因克

隆中最常用的方法。遗传病资源(家系资料)对于致病基因的克隆起重要作用。动物模型的建立也是致病基因克隆的重要手段。生物信息学是致病基因克隆的必需技术平台。成功鉴定一个致病基因往往需要遗传学、分子生物学、细胞生物学、生物化学、免疫学、病理学、临床医学等多学科间的密切配合。

定位候选克隆方法往往是从疾病家系分析或动物模型入手。根据系谱图确定某种遗传病基因的遗传方式,常通过全基因组扫描,分析 STR 或 SNP 单体型与疾病的连锁关系,确定关键区域(critical region)和进行精细遗传作图。

在缺乏家系的情况下,小鼠模型是定位克隆中最常用的。传统的定位候选克隆方法中,首先是建立小鼠回交系(backcross),通过遗传标志(genetic marker)与连锁性状(linkage trait)的分析,进行遗传作图。将致病基因定位于染色体某区段后,确定该致病基因所在的关键区。根据该区域的 STS 标志,制作物理图;并利用 STS 标志筛选阳性 BAC 克隆,通过对挑选的 BAC 克隆的末端测序、DNA 指纹图等建立 BAC contig,对关键 BAC 克隆,进行"鸟枪法"(shotgun)全长测序,或通过外显子捕获(exon trapping)、cDNA 筛选(cDNA selection)等方法,在关键 BAC 克隆中找出关键区内所含有的基因,从中筛选出候选基因,通过突变的检测和功能鉴定,最终找出人的同源基因,在患者中进行突变筛查,确定致病基因。

随着人类基因组计划的完成和生物信息学的应用,定位克隆过程已大大简化。传统方法中平均需要 3~5 年才能克隆的基因,现在一般只需 1~2 年时间。在确定致病基因关键区后,通过 3.6 节将要叙及的基因组分析方法,可直接得到相应的物理图和获得该区域的全序列,并得到已知的或预测的基因,从中筛选候选基因进行突变分析。省略了以往的物理作图、BAC contig 的建立和基因预测分析的过程,也不再使用外显子捕获等基因筛选技术。

图 3-3 示意利用定位候选克隆方法确定 Hermansky-Pudlak 综合征-7(HPS-7)的致病基因 *DTNBP*1 的遗传与物理作图[116]。作者首先在已确认的一种 HPS 小鼠 sandy (*sdy*)模型中,利用回交方法,建立与 PWK 种系小鼠的回交系。*sdy* 是来源于 DBA/2J 种系的自发常染色体隐性突变,并通过连锁分析初步定位于小鼠 13 号染色体上。这两种品系的小鼠在进化距离上相距较远,推测有较多的分子标记区别于两者。通过对 1 250 只回交系小鼠 20 个 13 号染色体 STR 分子标记的分析,将 *sdy* 基因定位于 D13Mit244 和 D13Mit267 之间的关键区,两者的遗传距离约为 2.2 cM。利用这些 STR 分子标记作探针,筛选 RPCI-23 BAC 文库,建立了含 9 个 BAC 的 contig。同时利用基因组分析,确定了 *sdy* 基因关键区的物理图,定位于 13 号染色体 43 Mb 附近约 570 kb 的区域。该区域仅含两个已知基因 *jmj* 和 *Dtnbp*1。突变分析表明,在 *sdy/sdy* 小鼠中 *Dtnbp*1 基因的第 6,7 外显子缺失,无法检测到相应的 dysbindin 蛋白。利用 BAC 原核注射得到的转基因小鼠则能表达正常的 dysbindin 蛋白,并纠正其表型的改变。因此,确认 *Dtnbp*1 基因的突变是导致 sandy 小鼠的原因。然后在 HPS 患者中,对其同源基因 *DTNBP*1 进行突变筛查,也发现病理性突变,成为继 HPS1~HPS6 后发现的又一种 HPS 基因。

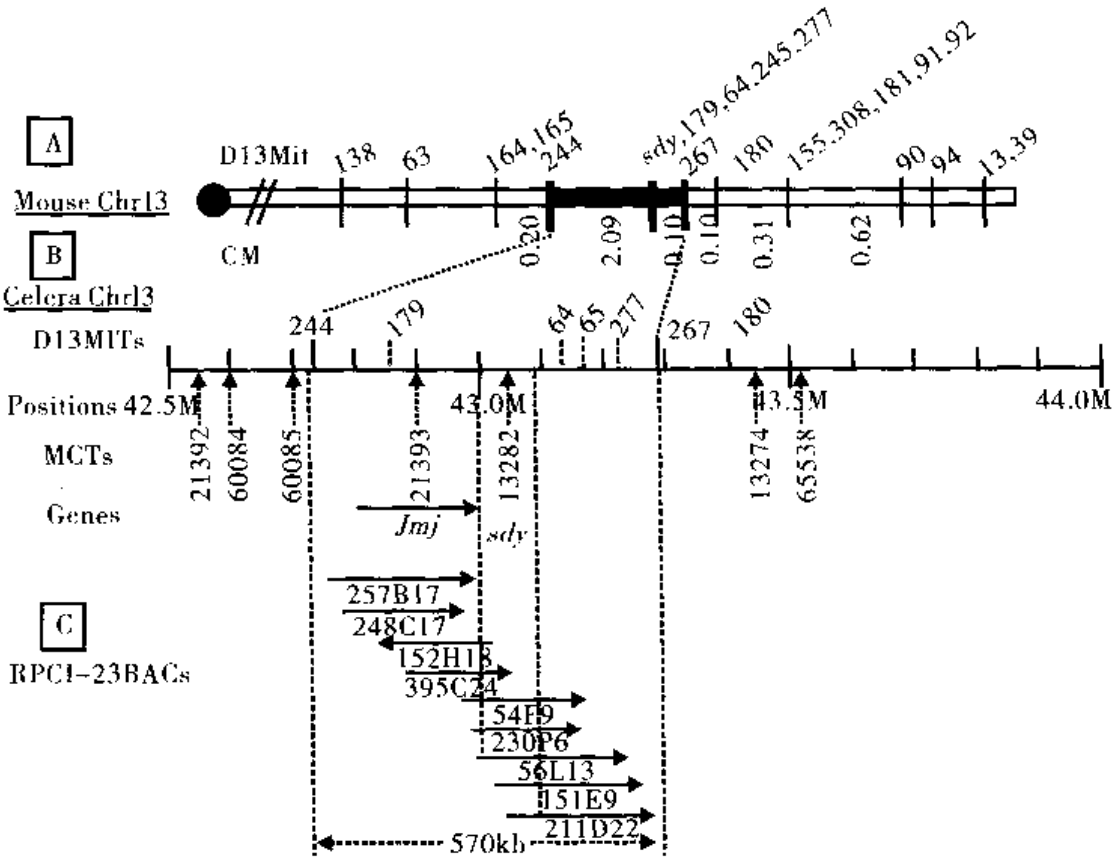


图 3-3 小鼠 *sdy* 基因的定位候选克隆——遗传作图与物理作图

3.6 基因组的组装与分析[17-19]

3.6.1 基因组图谱的整合

三大基因组数据库的组装所采用的信息管道(pipeline)有所不同,但都是根据国际人类基因组协作组的测序资料而进行。由各测序中心向 GenBank 存放高通量基因组序列(HTGS),建立相应的层叠路径文件(tiling path file, TPF),其中包括各 BAC 克隆的排列顺序、染色体位置、克隆相互重叠关系,以及缺口(gap)位置等。

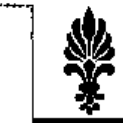
以下以 NCBI 人类基因组数据库为例,说明其组装方法。分以下几个过程。

3.6.1.1 输入数据的准备

测序资料首先需要用 MegaBLAST 程序对污染序列进行筛选,然后用 Repeat Masker 对重复序列进行标记,重复序列变为小写,再用 BLAST 检索已组装好的基因组如面包酵母、线虫、果蝇等,另外需要利用 Repeat Masker 对小鼠重复序列进行筛选,并用电子 PCR(e-PCR)筛选小鼠的 STS。任何序列若含有来自非人类的其他基因组序列,该序列即被删除。最后得到的可用序列是被缺口分隔的序列,其中的重复序列已被封闭好。

3.6.1.2 序列排版(sequence layout)

此步骤主要是在基因组 contig 基础上进行。首先是对单个 BAC 克隆进行排版,这是根



据上述 TPF 文件和克隆重叠信息,结合染色体定位资料进行排列。利用 MegaBLAST 检索到的某些非 TPF 序列如果能向缺口区延伸,此时也被拼装起来。然后对相互重叠的 BAC 进行拼接,去除冗余(redundant)序列。位于不同染色体但也相互重叠的 BAC 不被考虑拼接,完全被包含在另一更大 BAC 克隆的序列被删除。经此步骤得到的序列通常是来源于多个 BAC 的序列。

3.6.1.3 序列组装(sequence assembly)

在上述 contig 排版基础上,进一步检测相互重叠序列,再进行拼接,得到更长的合并序列(meld),这种合并序列的排列顺序还基于已知的 EST 或 mRNA 序列。两个不相连的合并序列之间的缺口用 100 个"N"填充。最后根据已知的 STS 标记,BAC 末端序列等将这些合并序列排列到染色体相应位置,称之为 scaffold。各 scaffold 排布于染色体上,即形成该染色体的序列草图(draft sequence map)。

3.6.1.4 基因预测

序列图拼装后的一大任务是进行基因预测。已有不同的基因预测软件,较常用的有 Genscan, Fgenesh + + 等。这些软件都是建立在已知的基因结构特征上,同时利用已知基因进行校验,准确性较高,但仍有一定比率的假阳性和假阴性。预测的基因如有对应的 EST 相吻合,常作为该预测基因存在的佐证。

3.6.1.5 图谱特征(feature)的整合

序列图组装后,通常需要将有关的序列变异(SNP)、STS 标记、BAC 克隆的 FISH 定位、已知和预测基因定位及其基因结构模型、已知 EST 或 mRNA 序列、蛋白质序列、与小鼠和大鼠的同源序列等整合到序列图的相应位置,最后形成 NCBI 的 Map Viewer 数据库。BAC 克隆和 STS 的排列通常用 MegaBLAST 和 e - PCR 分析法。SNP 的排列用 BLAST 法。已知基因及其相关的 mRNA、蛋白质序列的排列是根据 RefSeq 数据库,利用 MegaBLAST 方法进行。选择性拼接形成的 mRNA 也排列在同一基因位置。这些相关序列的排列一般要求与基因组序列的相同性在 95% 以上。一个容易出现的误差通常发生在 mRNA 与对应的基因组序列出现差异的情况,这可能是测序错误、组装错误、多态位点或者同系基因,计算机对此无法自动纠正。

3.6.1.6 建立相关链接(LocusLink)

用于佐证已知基因或预测基因的相关 RefSeq、mRNA 或 EST、蛋白质序列等可通过相互链接进行查找。其他与 Map Viewer 链接的常用数据库包括 dbSNP, UniSTS, UniGene, GenBank, COG, PubMed, BLAST 等。

UCSC 和 Ensembl 数据库的整合方法大体与 NCBI 相似,但图谱中所包含的特征有所不同,建议读者使用时根据个人喜好,选择某一数据库为主,同时熟悉其他数据库的功能,相互取长补短,灵活使用,久而久之便形成自己的一套检索习惯。NCBI 数据库是三大数据库中信息量最大的,相互链接较容易。但其他两大数据库也有其优点,例如在检索与 mRNA 对应的基因组序列时,UCSC 的 BLAT search 功能较 NCBI 的 BLAST search 便利。UCSC 和 Ensembl 的图谱较 NCBI 更直观,特别是 UCSC 图谱中有对应位置的暗画线,较准确反应基因的定位及其邻近关系。

3.6.2 基因在基因组中的定位方法

以下分别叙述如何在 NCBI、UCSC、Ensembl 三大数据库中定位某个基因的方法。

3.6.2.1 在 NCBI 中的定位

进入 NCBI 主页，点击进入 map viewer，在 search 下拉菜单中选择需检索的基因组，在 search for 框中键入基因名称，点击 go。显示结果中上方为该基因在基因组特定染色体上的定位图，若为同系基因，将显示多个不同位点。下方为查找结果的文字描述，分别列出染色体号、基因名称、图谱成分、标记类型和图谱类型。图谱类型中，人类基因组图分为细胞遗传图和序列图两种，可分别点击进入，得到相应的定位资料。若点击进入该条目中的图谱成分（map element），会得到该基因的精细定位图谱。点击右上方的 Maps & Options，可按用户要求选择所需图谱，如常用的细胞遗传图（Genes_Cyto）、基因序列图（Genes_seq）。在基因序列图中，用线条标明该基因的跨度和外显子数目和大致位置，同时有一系列链接点。sv 表示序列查看，ev 表示该基因序列的支持证据，hm 表示人与小鼠同源基因图，seq 为该基因的全部基因组序列，mm 表示基因结构组成。在定位图谱的左侧蓝色边框中可选择放大（zoom out）或缩小（zoom in）该观察区域，以便于了解该基因的近邻关系。

3.6.2.2 在 UCSC 中的定位

进入 UCSC 主页，在左边框上方点击 Browser 后，在物种下拉菜单中选择所需分析的基因组，如人类基因组，然后选择序列版本，缺损值为最新版本。在 position 框中键入基因符号后，点击 submit。结果显示已知基因和 RNA 相关序列。在已知基因栏目中找到所需查找的条目，点击进入。显示结果上方为基因定位图谱，下方为显示内容参数选择。如只需查找该基因的染色体定位和序列图，参数选择如下：Base Position 选 on，Chromosome Band 选 dense，Known Genes 选 full，RefSeq 选 full，其余参数均选 hide。修改参数选项后，需点击参数表右上方的 refresh 才能执行。在图谱上方 position 框中自动显示的数据即为该基因的序列图位置。如需放大或缩小观察该基因的图谱位置，可在图谱最上方参数中点击 zoom out 或 zoom in 到所需倍数，以了解其近邻关系。若需查找该基因的有关序列，在图谱左侧相应已知基因或 RefSeq 栏目中点击该代码，进入该基因的信息窗，通过 Links to sequence 各链接点，可分别得到该基因的蛋白质、mRNA 和基因组序列等。

3.6.2.3 在 Ensembl 中的定位

Ensembl 数据库中收录了人、小鼠、斑纹鱼、蚊虫等基因组。例如选择人类基因组，进入其主菜单。在 search 下拉菜单中选择 Gene，然后在 indexes for 框中键入所查找的基因符号，点击 Lookup，则得到所需结果。显示基因条目都有一个 Ensembl 编号，以 ENSG 符号开始，后接一串数据，组成该基因的 ID 号。点击该 ID，进入该基因的详细记录文字信息。其中包括与相应蛋白质结构域和家族的链接点。在结果第二栏 Genomic Location 中，显示该基因的序列图位置和染色体编号。点击该位置，进入 Contig View 窗，上方图谱为概要图，显示该基因在染色体的定位基本信息，红色框内为该基因的位置，显示内容包括染色体带、DNA contig、分子标记和基因图。下方为详细图，只显示所查找基因所在区域的相关信息。在 Features 下拉菜单中可进行选择观察所需内容，点击 Refresh 更新选项。进一步可通过缩、放选择观察该区域及其与邻近基因的关系。若返回前面总的基因查找显示结果菜单，最后一行

是一链接点，进入该基因的结构描述窗，分别显示该基因的 cDNA 序列及其在序列图中的定位、外显子－内含子连接区位置、各外显子序列及其大小等信息。

值得注意的是，这三大数据库中，对于同一基因有时出现不同的基因命名，甚至序列有细微差别（可用 NCBI 的 BLAST 2 sequence 鉴别其异同），特别是那些预测的基因。这时横向比较该基因在不同数据库的定位时有些困难，需要借助于邻近已知基因的关系来确认。同时各数据库所用的序列版本可能有所不同，同一基因在不同数据库的序列图定位也可能出现细微差别。一般地，描述某个基因的定位时，须加注所用数据库和序列版本，如本文采用的 UCSC 人类基因组版本是 2003 年 4 月版。最后一点需要牢记的是，这三大数据库在不断更新，以上所述的检索方法在新的版本中，可能有所不同，检索的结果也可能有些差异，希望读者能经常上网浏览，了解更新后版本的新特征。

3.6.3　基因的结构组成与序列检索

若已知基因的基本信息，需要检索其 cDNA，确定基因结构，得到基因组 DNA 和蛋白质序列，这可以任选 NCBI、UCSC、Ensembl 三大数据库之一进行检索。分以下几种不同情况。

（1）对于已知基因，可以查找 NCBI 的 UniGene，每个条目下已注明该基因的蛋白质、mRNA 和相关信息的链接点。点击相应部位可得到相应的序列。或者按 3.6.2 中所述基因查找方法，利用 UCSC 或 Ensembl 数据库也可查找得到基因的结构组成信息。通过 NCBI 主页点击进入 Reference sequence project，可查看到已知基因的 RefSeq 条目，这是 NCBI 的数据库维护人员通过文献资料、GenBank 条目和基因组分析资料等，对每个已知基因确定其参考 cDNA 序列。如图 3－4 所示，为 NCBI 数据库中检索到的小鼠 *Dtnbp*1 基因的 RefSeq 条目（NM_025772）。

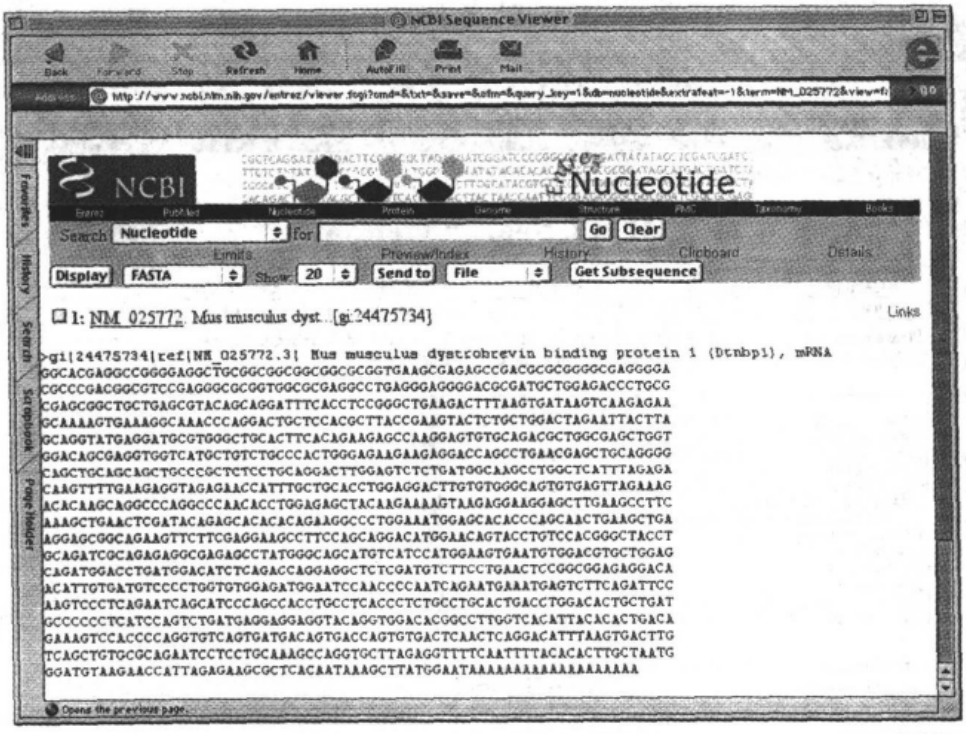


图 3－4　从 NCBI 的 RefSeq 中检索得到的小鼠 *Dtnbp*1（NM_025772）全长 cDNA 序列（FASTA 格式）

(2)如果只知道所需查找基因的一般信息,可以先从 NCBI 的 GenBank(或 nucleotide 选项)查找框中键入基因的相关名称,可以是全名,也可以是基因简写。从所显示的条目中选择含整个 ORF(基因条目中已注明 complete cds)的 cDNA 或 mRNA 序列;在 DISPLAY 下拉菜单中,先选择 summary 选项,可获得该基因条目的基本信息,这是标准的 GenBank 条目格式。其中已注明 ORF 的起始和终止核苷酸位置,并列出相应的编码蛋白质序列。点击蛋白质序列条目,可进入 GenBank 的蛋白著录格式。在 DISPLAY 菜单中选择 FASTA 格式(参见图 3 -4),将所得序列复制后,另外存为纯文本文件或 WORD 文件,也可同时打开 UCSC 数据库,进入 BLAT search 选项,直接将序列复制、粘贴到查找框中(参见图 3 -5),或在序列框下面小框中输入已保存序列的文件名。选择所需查找的是人、小鼠、大鼠或 SARS 基因组等;序列版本(freeze)的缺损值为该基因组的最新版本;分析对象(query type)为 DNA(也可输入蛋白质序列,此时选择分析对象为 protein,该缺损值为 guess);点击 submit。所显示的查找结果为该基因组中相匹配的基因序列或同系基因序列,通常完全相同或相似性最好的序列排在首位(参见图 3 -6)。根据序列大小、染色体定位等确认所需条目,点击该条目左侧的 details 进入 alignment 对话框(参见图 3 -7)。左边框显示的 block 数目一般为外显子数目。由上而下,分别为 cDNA、genomic DNA 和序列比对结果。显示的 cDNA 与所输入的序列不同之处,或无匹配的 DNA 序列时,用小写标明,两外显子相连接处的两个核苷酸用浅色标明,5'和 3'非编码区用红色标明。基因组序列中,外显子用大写,内含子用小写,同时注明染色体号及其相应核苷酸位置。序列比对具体显示所输入序列与基因组中序列的异同。图 3 -5、图 3 -6、图 3 -7 分别示意将图 3 -4 的小鼠 *Dtnbp*1 基因 cDNA 序列复制、粘贴后查找其相应基因结构的过程。

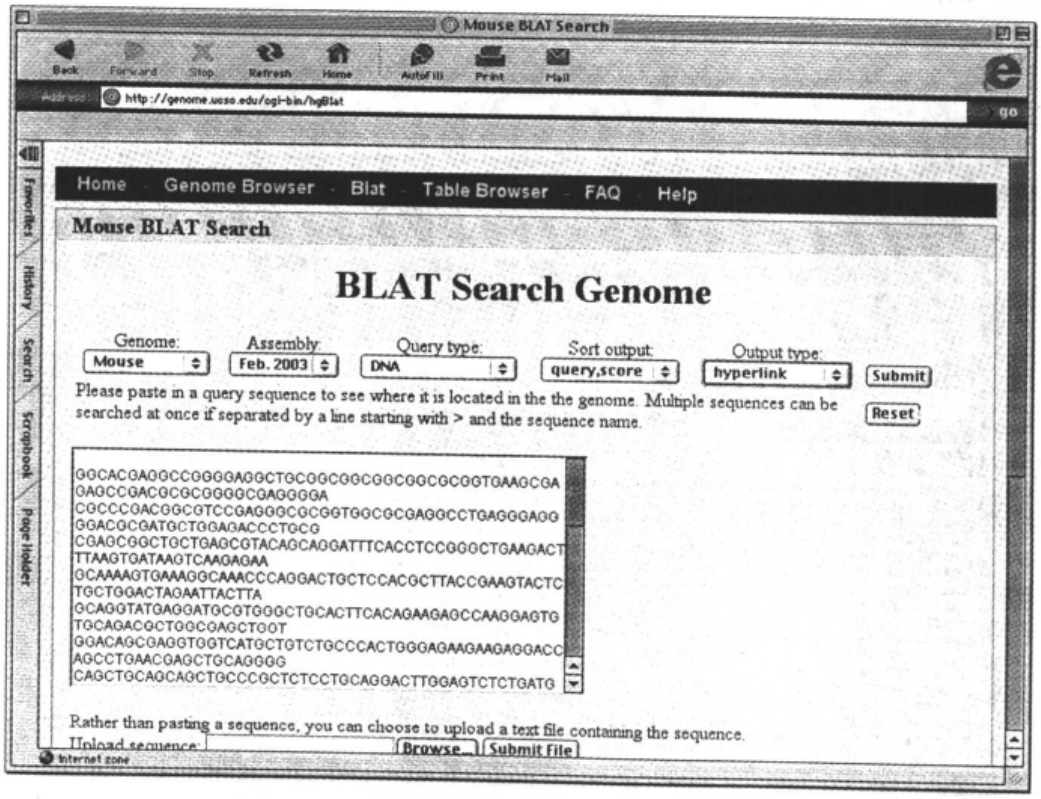


图 3 -5　将图 3 -4 中的序列粘贴到 BLAT 检索框中进行检索

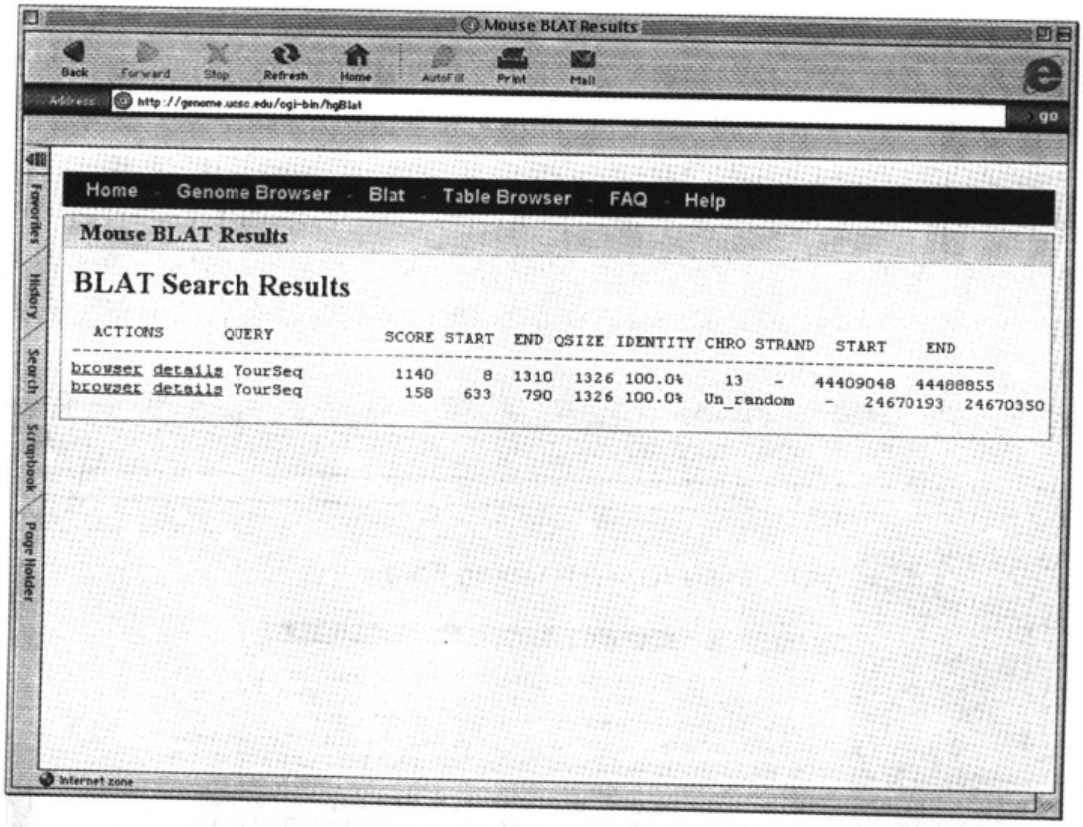


图3－6　点击图3－5中submit得到的BLAT检索结果

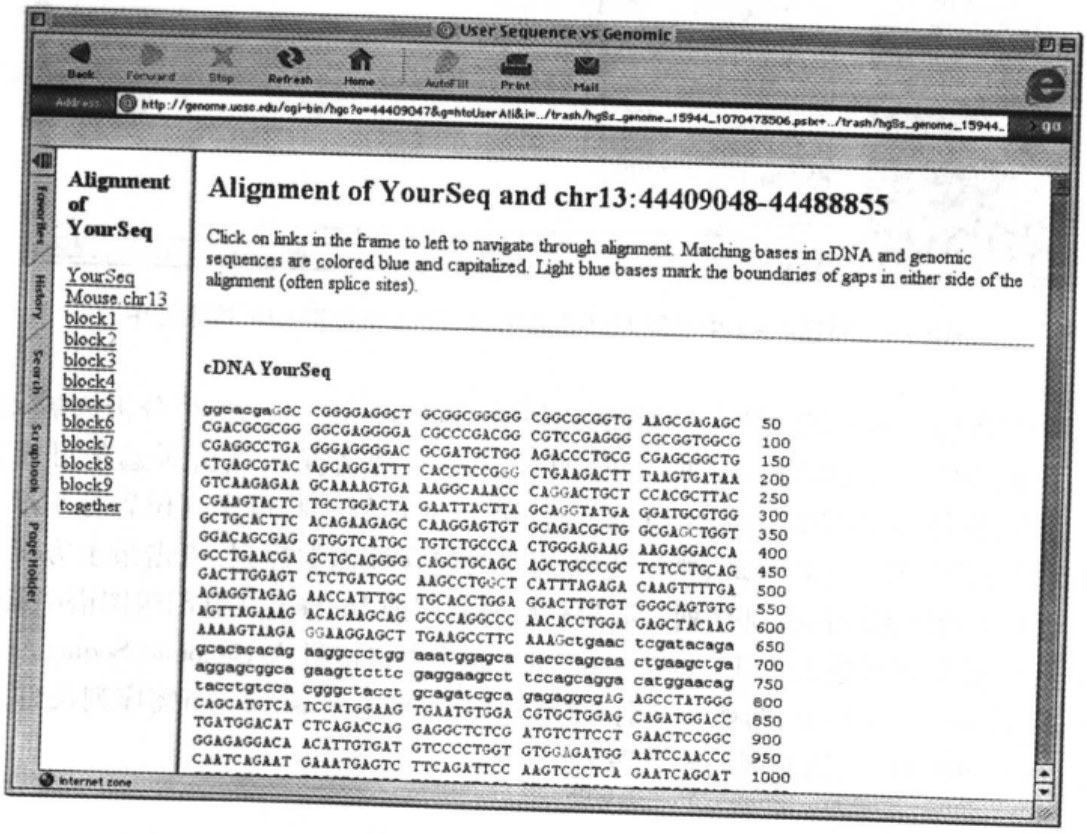


图3－7　点击图3－6结果显示栏中第一行的details得到的小鼠 *Dtnbp1* 基因结构组成与序列

(3)如果只是通过测序获得一个 cDNA 序列,须查找对应的基因结构,可按上述第二种情况中的 UCSC 查找方式,将序列粘贴到查找框中,获得基因组结构组成信息。在 GenBank 中检索某个基因时,常可同时获得该基因的 mRNA 和基因组序列,需要了解两者的对应关系时,可选用 NCBI 的 Spidey 工具(http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey)进行排列。得到该基因组序列后,进入 NCBI 主页,点击右侧 ORF Finder,将 cDNA 序列粘贴到查找框中(参见图 3-8),结果显示不同的阅读框。点击最长的阅读框(参见图 3-9),得到相应的编码子序列,确认是否具有完整的 ORF,即同时具有启动子和终止密码。这样可得到所编码蛋白质的序列。

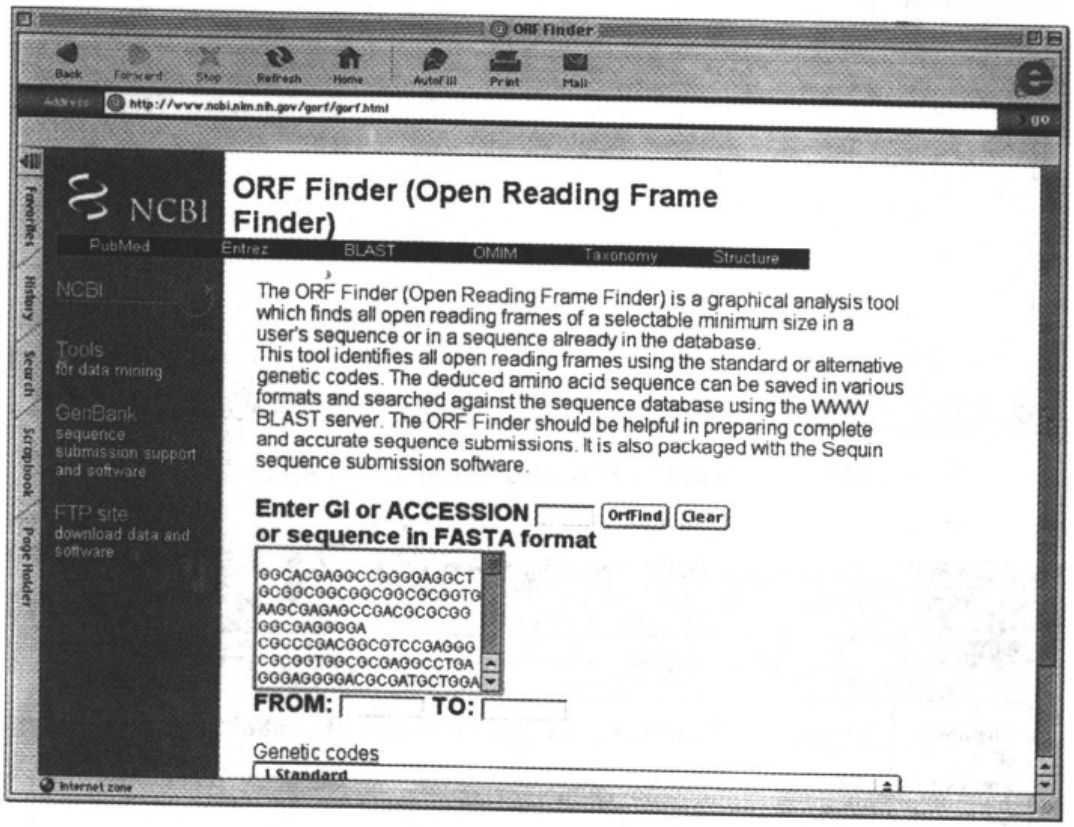


图 3-8　将图 3-4 中小鼠 *Dtnbp*1 基因的 cDNA 粘贴到 ORF 检索框中

(4)如果需要检索该基因的上、下游侧翼序列(flanking sequence),以分析该基因潜在的调节区,可在 UCSC 的 BLAT search 结果中(上述第二种情况),点击所选条目最左侧的 browser 进入基因组整合图谱,在图谱上方的 position 对话框中将原序列位置分别增加所需包括的侧翼序列长短,点击 jump,则所更新的图谱包含该侧翼序列,点击最上方工具条中 DNA,即可获得所需基因的全部序列。如果是已知基因或 RefSeq,在基因组图谱中点击该基因名称(如 3.6.2 节所述),在 Links to sequence 选项中点击进入 Genomic Sequence,在 Get Genomic Sequence Near Gene 选项中分别选择 Upstream 和 Downstream 所需序列长短,点击最下方的 submit,即可得到该基因的侧翼序列。

3.6.4　定位克隆中关键区的定位和候选基因的筛选

利用 3.5 节所述方法将疾病基因定位在某个关键区后,接下来的关键问题是如何通过

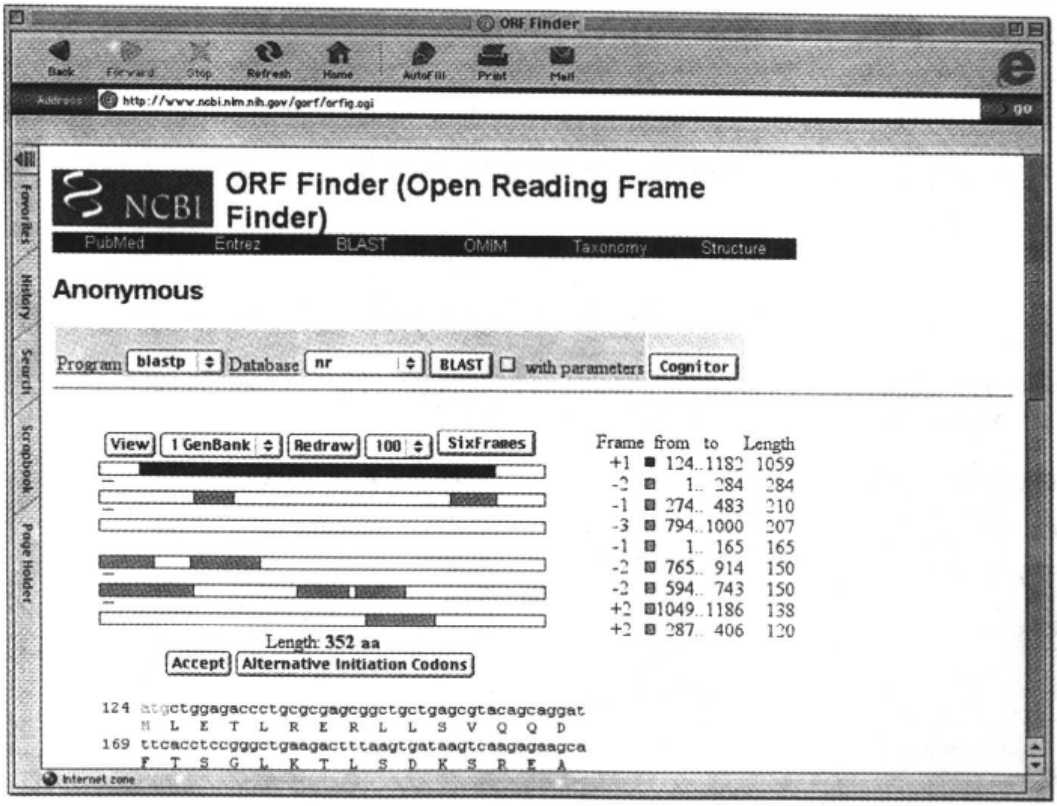


图3－9　点击图3－8的ORF Finder的结果和点击最长的ORF所得结果

基因组分析确定候选基因。关键区的两端为STS分子标记。下面分别以UCSC和NCBI数据库为例，说明其候选基因的定位与筛选方法。

进入UCSC主页后，点击左侧Browser，选择所需分析的基因组，如人类基因组，在查找框内键入两端的STS标记，如“D10S1676；D10S1675”，两个STS间以分号隔开，按submit后，显示图谱为该区域的整合图，重新设定参数表。Base Position/on，Chromosome Band/dense，STS markers/full，Known Genes/full，RefSeq/full，Ensembl Genes/dense，Fgenesh＋＋Genes/dense，Genscan Genes/dense，Human mRNA/full，Spliced EST/dense，Human EST/dense，其余参数选择均为hide，按refresh。所得到的图谱中主要显示各STS位置，所包括的已知基因和对应的mRNA或EST。需要显示预测的基因，点击图谱中相应预测软件所在位置，如Genscan，Fgenesh＋＋。如需进入某个已知基因、预测基因或EST的相关信息和序列，将鼠标移到图谱左侧的名称上，点击进入。列出所有已知基因和预测基因条目，即为该关键区内的候选基因。筛选时一般从已知基因开始，首选那些功能相关的基因。但应注意，一些较大的已知基因可能具有不同的结构域，因而表现为一个基因多种功能，不能武断地排除某些已知功能的基因。如果已知基因中未发现突变，则需要筛查那些预测基因，通常先从那些有EST佐证的预测基因开始，逐个排查。

若选择NCBI数据库，首先从主页进入需查找的基因组，如人类基因组，在染色体图谱中，点击进入该基因所在的染色体，显示的图谱为缺损项设置。点击右上方的Maps & Options链接点，首先在右上方的Region Shown两个方格中键入关键区两侧的STS名称，然后进行添加或删除选项。添加时，在左框中选择参数后，按ADD，则进入右框；删除时，在右框

选择参数，按 REMOVE(参见图 3－10)。一般至少需要选择 UniSTS，Gene 两个选项。下方页面大小(Page Length)的缺损值为 20 条，可根据需要改变，要求能显示所有的条目。各参数选择完毕后，按 Apply，则显示所需要的图谱。显示的基因图中，包括所有已知和预测的基因，通过相关链接点，可查看每个基因条目的相关信息和序列(参见图 3－11)。按前面所述方法逐步筛选候选基因。

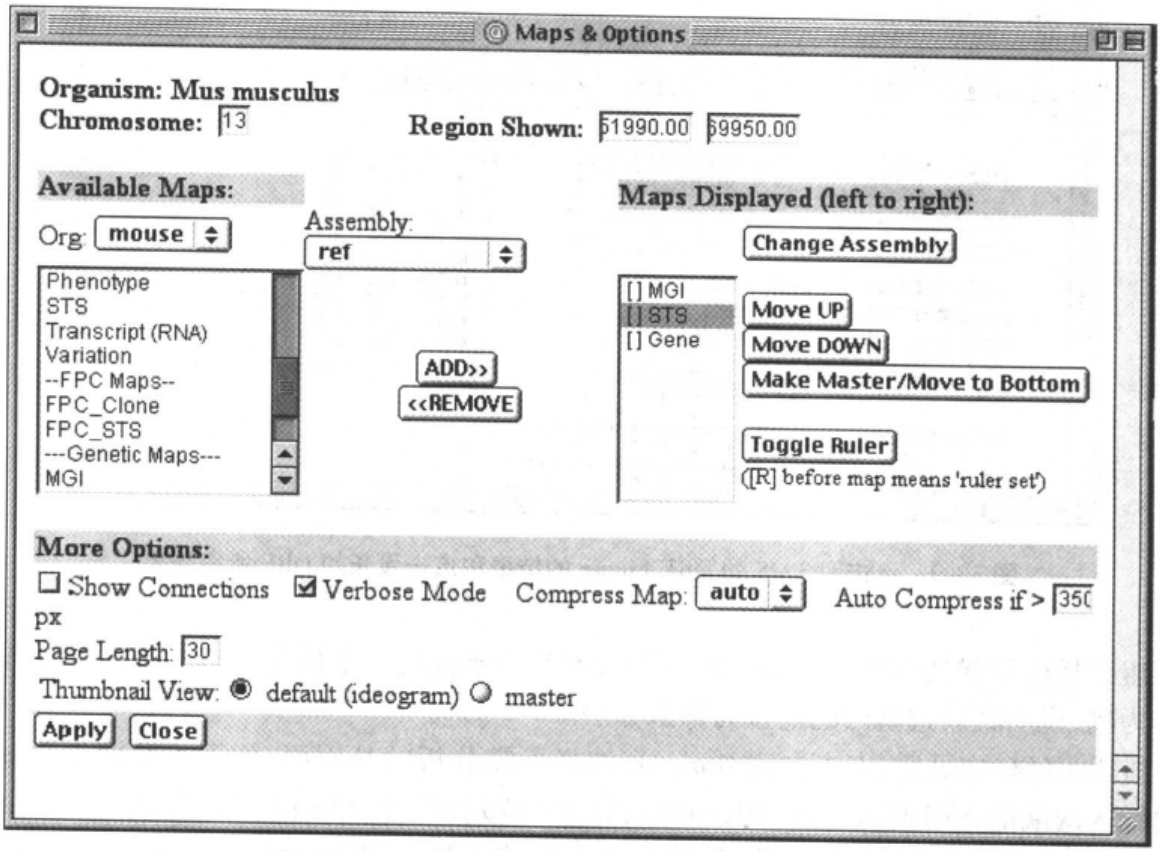


图 3－10　通过 NCBI 查找图 3－3 中 *sdy* 基因关键区内的候选基因(关于图谱的选项)

3.6.5　同系基因和同源基因检索

对于同系基因或同源基因的检索，简便的方法是利用 UCSC 的 BLAT search 检索工具。将已知基因名称或序列按 3.6.2 节和 3.6.3 节所述进入 BLAT 查找框。对于同系基因选择与检索基因相同的基因组，检索所得结果，若有同系基因存在，则根据相同程度由高到低排列，点击相关链接点，可知该同系基因的定位资料和序列比对情况。如为同源基因查找，则选择与所查找基因不同的基因组，输入相应序列，进行 BLAT search，可得到同源基因的定位资料和序列比对情况。如需同时查找多个不同基因组的同源基因，则建议用 NCBI 的 BLAST search 工具进行检索。最近 NCBI 增加了 HomoloGene 的检索功能，可输入基因名称，直接查找其同源基因。

因限于篇幅，未能将上述各步检索结果的插图全部附上，建议读者在阅读本节时，先将这三大数据库网址同时打开，根据各自的工作领域，设定所需分析的基因，对照检索，便于理

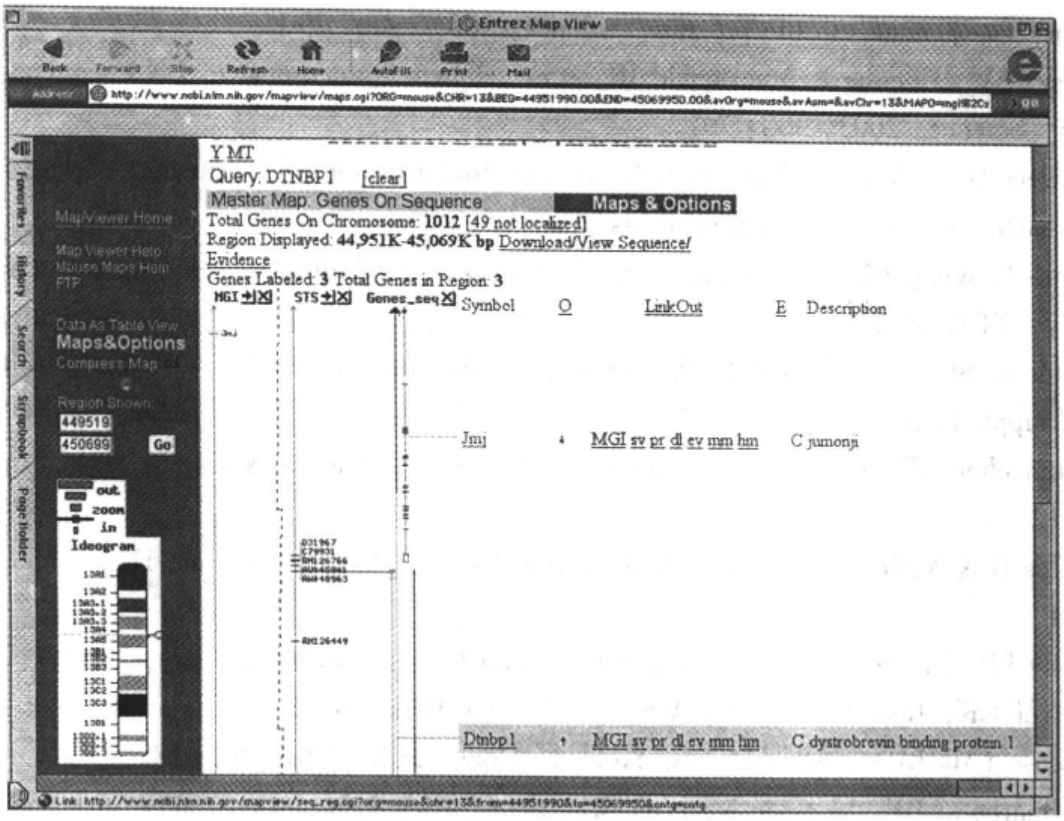


图3－11　根据图3－10选项所得到的遗传图、分子标记物理图及关键区内的候选基因

解和掌握。前面已经提到，基因组的分析是每位生物信息学工作者的必备知识，希望读者在以上所述方法的基础上，反复练习，灵活自如地应用。在掌握基因组分析方法的基础上，进一步参考本书第5、6章内容，进行有关DNA或蛋白质序列或结构的分析（备注：有关基因组医学中的常用术语参见http://www.genome.gov/glossary.cfm）。

（李　巍）

参　考　文　献

1　李巍，何蕴韶. 遗传咨询. 郑州：郑州大学出版社，2003.4～7

2　Collins FS, Morgan M, Patrinos A. The human genome project: Lessons from large－scale biology. Science, 2003, 300: 286

3　Collins FS, Green ED, Guttmacher AE, *et al*. A vision for the future of genomics research: A blueprint for the genomic era. Nature, 2003, 422: 835

4　The Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860

5　Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. Science, 2001, 291: 1 304

6　Frazier ME, Johnson GM, Thomassen DG, *et al*. Realizing the potential of the genome revo-

lution: The genomes to life program. Science, 2003, 300: 290

7 Scherer SW, Cheung J, MacDonald JR, *et al*. Human chromosome 7: DNA sequence and biology. Science, 2003, 300: 767

8 Krawetz SA, Womble DD. Introduction to Bioinformatics: A Theoretical and Practical Approach. New Jersey: Humana Press, 2003. 151 ~ 184

9 Licinio J, Wong ML. Pharmacogenomics: The Search for Individualized Therapies. Weiheim: Wiley - VCH, 2002. 9 ~ 82

10 Patterson SD, Aebersold RH. Proteomics: the first decade and beyond. Nat Genet, 2003, 33(suppl.): 311

11 Guttmacher AE, Collins FS. Genomic medicine - A primer. N Engl J Med, 2002, 347: 1 512

12 Childs B & Valle D. Genetics, biology and disease. Ann Rev Genom Hum Genet, 2000, 1: 1

13 Reilly PR. Public concern about genetics. Ann Rev Genom Hum Genet, 2000, 1: 485

14 Scriver CR, Beaudet AL, Sly WS, *et al*. The Metabolic & Molecular Basis of Inherited Disease (MMBID). 8th ed. New York: McGraw - Hill, 2001. 193 ~ 287

15 Gudbjartsson DF, Jonasson K, Frigg ML, *et al*. Allegro, a new computer program for multipoint linkage analysis. Nat Genet, 2000, 25: 12

16 Li W, Zhang Q, Oiso N, *et al*. Hermansky - Pudlak syndrome - 7 (HPS - 7) is caused by mutant dysbindin, a new member of the BLOC - 1 complex. Nat Genet, 2003, 35: 84

17 Wolfsberg TG, Wetterstrand KA, Guyer MS, *et al*. A user's guide to the human genome. Nat Genet, 2002, 32(suppl): 4

18 Baxevanis AD, Quellette BFF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2nd ed. New York: John Wiley & Sons, 2001. 94 ~ 105

19 Lesk AM. Introduction to Bioinformatics. Oxford: Oxford University Press, 2002. 62 ~ 114

第4章　生物信息学常用数据库介绍

生物信息学最初的定义是建立数据库，用于储存和分析生物学相关的信息。尽管近年来生物信息学的发展如此迅猛，赋予了生物信息学更丰富的含义，但数据库仍是生物信息学的最基本的内容，也成为生物信息学的重要分支——数据库生物信息学（database bioinformatics）。生物信息学中涉及的分子生物学数据库类型很多，常用的包括序列数据库、定位数据库、医学信息数据库、基因组数据库、结构数据库等，其中基因组数据库及其应用在第3章已作介绍，结构数据库将在第6章作详细介绍。有关数据度结构和管理系统的基本概念属于计算机科学的基础内容，这里不展开讨论。

4.1　分子生物学数据库

4.1.1　分子生物学数据库发展简史

分子生物学数据库在现代生命科学中起着非常重要的作用。作为生物信息学和基因组学中的关键内容，这些数据库主要解决DNA和蛋白质序列分析、功能基因组学、蛋白组学和生物大分子结构分析中的核心问题，从而为数据储存、查找和分发提供有效的途径。

最早的生物数据库是有关生物种类和性状方面，而不是我们现在所想到的有关序列方面。物理性状的遗传首先取决于其物理定位，然后是基因产物的存在，最终是由其序列中的遗传信息所决定。首个分子生物学数据库是蛋白质数据库，其动因源自Fred Sanger等于1955年对胰岛素一级结构的测定。此后Sanger的测定方法被广泛用于蛋白质生物化学研究，并得以对数以千计的蛋白质序列进行测定，用于阐明许多结构蛋白和代谢途径中的酶类特性，以及分子进化分析等。

早在20世纪60年代初期，Margaret Dayhoff就提出这样一个设想，即将文献中所有的蛋白质序列汇总在一起，可能是研究中非常有用的

工具。于是她和同事一道收集所有已知的序列并在一起发表,称之为"蛋白质序列及结构图谱"[1]。随后,每当新出现一种蛋白质序列,即与该图谱比对,以找出与其他蛋白质的关系,这就使得一些不同蛋白质之间序列相似区域被鉴定出来。Dayhoff 因此提出蛋白质超级家族的概念:即许多具有不同酶活性的蛋白质之间相互关联且来自于共同祖先。随着该图谱中序列的不断增多以及计算机的出现,数据库开始变成数字化文件,并可通过调制调解器(modem)以 300 波特的速率传送(1980 年)。到了 1983 年,美国 NIH 资助建立了该图谱的连续版,称之为蛋白质信息资源库(protein information resource, PIR)。数年后,日内瓦大学与欧洲分子生物学实验室(EMBL)合作建立了著名的 SWISS - PROT 数据库,该数据库包含 PIR 的所有信息,同时含有丰富的注释,并可链接到其他数据库。近年来 SWISS - PROT 数据库已成为蛋白质数据库的标准。

接下来的序列信息大爆炸时代的到来始于 EMBL 的 DNA 序列数据库及随后于 1982 年建立的 GenBank。GenBank 最初是由美国 Los Alamos 国立实验室建立的。1988 年,美国组建了国家生物技术信息中心(National Center for Biotechnology Information, NCBI),并正式接管了 GenBank。在 20 世纪 80 年代早期,只有几个主要的数据库,1982 年存放于 GenBank 中的序列也仅有 606 条共 680 338 个核苷酸序列。当时没有人会预料到数据库会像今天这样如此庞大。那时的数据库主要靠用户自己管理,而不是集中管理,数据库中的信息由存放者自己修改。序列的注释也全由提交者负责,更新速度慢,且出现许多冗余序列。1988 年后,GenBank 中的数据呈指数级增长。到了 2003 年 2 月,GenBank 中已经有超过 2 200 万条序列,包含 280 亿个核苷酸。这些序列每日都在更新,并每两个月通过 Internet 完全发布出来。

分子生物学数据库的第三次浪潮的到来起源于各类物种的基因组数据库。自 1995 年两种细菌(流感嗜血杆菌和解脲支原体)的基因组全序列被破译以来[2],迄今已报道 400 多种基因组序列。基因组数据库的建立是为了满足各种不同的数据储存需要如遗传图、物理图、核苷酸和氨基酸序列等(参见第 3 章)。这些数据库代表不同分辨率的基因组结构和不同层面的基因组功能,是根据生物学分类而排列,而不是按书目编排体例。

4.1.2 分子生物学数据库分类

分子生物学数据库中的信息可以是 DNA 序列,保守的 DNA 结构域、基因组、基因表达、蛋白质序列、蛋白质家族、基因突变、基因多态性和代谢途径。目前已有许多数据库,通过 Internet 在搜索引擎可查到数百个数据库。一般而言,分子生物学数据库的分类依照数据类型、物种类型、数据来源、进入便利性、数据录入方式和侧重点等。

最先使用也是最流行的数据库是基于数据类型如 DNA、RNA、EST、蛋白质而建立的。例如,NCBI 中的 UniGene 数据库(http://www.ncbi.nlm.nih.gov/UniGene)是一种 DNA 序列数据库。该数据库是基于实验数据从 GenBank 中自动分出来的一种根据基因类别而建立的非冗余性数据库。每个 UniGene 条目代表一种独特的基因序列及其相关信息,如组织表达类型和基因定位等。除了已明确的基因外,该数据库还包括成千上万的新 EST(expressed sequence tag)序列。UniGene 提供了转录组水平的自动分析系统,解决了困扰许多研究者的转录序列中的冗余(redundancy)现象,以及常见的实验差错问题。因此,*UniGene* 可作为新基因发现或其他大规模研究计划如基因定位、基因表达、全长 cDNA 研究项目的重

要资源。另外，无数基因克隆的研究项目也从该数据库中受益非浅。

数据库的第二种分类方法是根据物种类型。近年来物种数据库显著增加用于储存该基因组中有关结构和功能基因组信息。一些主要的物种数据库如人类基因组的 GDB(http://gdbwww.gdb.org/) 和 NCBI(http://www.ncbi.nlm.nih.gov/genome/seq/)数据库，果蝇的 BDGP(http://www.fruitfly.org/)和 FlyBase(http://flybase.harvard.edu:7081/)数据库，酵母的 SGD(http://www.yeastgenome.org/)和 CYPD(http://mips.gsf.de/genre/proj/yeast/index.jsp)数据库，线虫的 Wormbase 数据库(http://www.wormbase.org/)，大肠杆菌数据库(http://www.genome.wisc.edu/)，拟南芥(Arabidopsis)数据库(http://www.arabidopsis.org/)等。

所有这些物种数据库包含各种不同的相应物种的信息，其数据库内容可有不同来源。例如，TAIR 数据库，其前身为拟南芥数据库(AtDB)[3]，是由美国自然科学基金资助的合作项目，收集、修正和发布有关拟南芥的信息。TAIR 提供有关基因、分子标记、多态性、图谱、序列、克隆、DNA 和库存种子、基因和蛋白家族等多种信息。此外，用户还能在该数据库中找到拟南芥研究人员所发表的相关论文。目前，该数据库已有 40 万条核苷酸序列、20 万个突变株、4 千个分子标记、9 万个多态位点、600 多个微阵列实验结果，以及多达 1 万 4 千余篇参考文献[4]。自 TAIR 公布以来(http://arabidopsis.org/usage/monthly/2002/05/01/index.html)，其月访问用户量由 1999 年 11 月的 1 万 3 千人次，到 2003 年 7 月已增加到约 5 万 3 千人次。

第三类分子生物学数据库分类是根据数据来源，通常分为两个业类：原始数据库和衍生数据库。原始数据库一般只存储递交者寄送的序列，例如 GenBank 和 PDB 等。衍生数据库是根据原始数据库编译得到的。数据库中的资料或者经过人工校对(curation)、或者运行程序自动编辑。人工校对的数据通常质量很高但耗费较高、且更新较慢，如 NCBI 中的 RefSeq。而自动产生的数据库耗费低、更新快，但与人工方法相比，质量较差，如 COGs 和 TrEMBL。TrEMBL 是对 SWISS - PROT 数据库的补充，以解决该数据库在基因组计划中输入到序列数据库时，人工编辑所面临的耗时长、工作量大的问题。维持 SWISS - PROT 数据库的高质量是必要的，同时尽快更新序列也很重要[5]。TrEMBL 产生于 1996 年，根据 DNA 序列利用计算机注释各条目。这些条目来自核苷酸数据库中的编码序列(coding sequence, CDS)和文献中的蛋白质序列，以及用户直接递交的蛋白质序列，但不包括 SWISS - PROT 中已经存在的编码序列。TrEMBL 对 EMBL 中的核苷酸序列每周更新一次。24.6 版本的 TrEMBL 包含 94 万个条目，含 291 兆的氨基酸。数据库依据功能或种系发生进行分割，如病毒、灵长类或细菌数据库，可看做衍生数据库。

数据库的侧重点(focus)是分类的另一标准。不管是原始数据库、衍生数据库还是物种数据库，每个数据库都有其侧重点，例如序列数据库、图谱数据库、基因表达数据库和医学信息数据库等。如前所述，序列数据库主要用来存放 DNA 或蛋白质的序列信息。大型序列数据库包括 NCBI 的 GenBank、EBI 的 EMBL、日本的 DDBJ 以及瑞士的 SWISS - PROT。图谱数据库包括定位、排列次序、分子标记之间的距离、克隆组集(contig)、探针、物理图等信息。基因表达数据库如 NCBI 的 GEO(http://www.ncbi.nlm.nih.gov/geo/)，收集基因表达和微阵列杂交数据并经过校对，可提供在线基因表达肿瘤浏览、查询和检索的有关资源。医学信

息数据库收集遗传性疾病及相关基因突变的信息，如 OMIM（http://www.ncbi.nlm.nih.gov/omim/）。该数据库是由 John Hopkins 大学 Victor A McKusick 及其同事等编辑的一种人类基因和遗传病目录。以下将详细讨论一些常见序列数据库、图谱数据库和医学数据库的建库历史、数据结构的分布、数据递交方式和查找方法等。

4.2 序列数据库

4.2.1 核苷酸序列数据库

核苷酸序列数据库(nucleotide sequence database)是最常见的一类序列数据库，已有 30 多种。一些数据库如 EMBL(http://www.ebi.ac.uk/embl/)和 GenBank（http://www.ncbi.nlm.nih.gov/Genbank/Genbank Search.html)含有各种物种的 DNA 序列。另一些数据库提供 DNA 的三维结构，如 NDB（http://ndbserver.rutgers.edu/），该数据库收集有关核酸的结构信息。还有一些数据库仅收集特殊的 DNA 或 RNA 序列信息，如 RDP（http://rdp.cme.msu.edu/html/），该数据库提供核糖体相关的数据服务，包括在线资料分析、rRNA 衍生的进化树、rRNA 序列比对和分析。PLACE（http://www.dna.affrc.go.jp/htdocs/PLACE/）是一种储存植物顺式作用调节性 DNA 元件的数据库。一些数据库收集有关分子或寡核苷酸探针，如 MPDB（http://www.biotech.ist.unige.it/interlab/mpdb.html）；并收集分子生物学中常用的 2 600 多种载体 DNA 序列，如 vectorDB(http://life.nthu.edu.tw/~g854202/Vecdtb.html)。

所有这些核苷酸数据库中，EMBL、DDJB 和 GenBank 是国际性核苷酸数据库的主干。这三大数据库之间每天都相互交换信息。用户只要向其中之一递交序列，且不管这些序列是否将被发表。这三大数据库依照相同的文献指南(The DDBJ/EMBL/GenBank Feature Table Definition)来规范数据库条目的内容和句法。该指南保证用同一种格式进行数据库之间的数据共享，并与现行的生物信息学软件兼容，反映分子生物学和普通生物学的进展情况。迄今为止，这 3 个数据库共包含 2 300 万个记录，其中 GenBank 占 71%，DDBJ 占 19.6%，EMBL 占 9.4%。

4.2.1.1 EMBL 核苷酸数据库

EMBL 是欧洲的主要核苷酸数据库资源，始建于 1980 年，将各个公共来源的核苷酸序列进行整合、整理和分类。该数据库由位于英国剑桥附近的欧洲生物信息学研究所(European Bioinformatics Institute, EBI)维护，并与 GenBank 和 DDBJ 共同协作。EMBL 数据库已处理自 1982 年以来全世界范围内所公布的序列。基因组计划的序列、各研究人员直接递交的序列以及由欧洲专利事务所发送的专利序列是该数据库的序列增加的主要来源。EMBL 数据库储存和维护是用 ORACLE 数据管理系统，可通过序列检索系统(SRS)、通过互联网进行查找[6]。

EMBL 数据库包括两个主要部分。第一部分是发行区(release section)(EMBLREL)，每隔 3 个月对外发行一次。第二部分是序列每日增添新区(new section)(EMBLNEW)。最新的 75 发行版含 25 214 767 个序列条目，包括 32 195 012 823 个核苷酸单位。EMBL 被划分

为19个组(division),分别反映不同的用户感兴趣的数据子集。这些组的区分原则是根据分类学(taxonomy)(如HUM代表人类,PLN代表植物,PRO代表原核生物等)。一些例外的是HTG(high throughput genome sequences)、GSS(genome survey sequences)和EST(expressed sequence tags)等,是根据某些资料特性进行分类。

EMBL数据库的结构是按flatfile(平面文件)格式进行布局的,该格式可被许多序列分析软件所支持,其数据库结构方式适合读者使用,见附录4-1。EMBL的flatfile格式包括一系列严格控制的行类型和四大主要数据区(block)。第一个区包括描述和标识符,如条目名称、保密状况、分子类型、分类、序列总长度(ID行)等基本描述内容;标识符有登记号码(AC)、序列版本(SV)、创建和更新日期(DT)、序列简单陈述(DE)、关键词(KW)、分类方法(OS, OC)以及相关数据库的条目(DR)等。第二个区是引文区,包括引文详细内容(RX, RA, RT和RL)以及原递交者姓名(RA)和联系方式(RL)。第三个区是由许多特征(FT)行组成,包括序列的特征,如详细来源、生物特征、特征性定位和限定词等。最后一个数据区由序列资料、长度和碱基组成(SQ)。

有几种不同的方法向EMBL递交数据。Webin是EMBL常用的递交媒体。Webin通过一系列交互式WWW表格指导用户发送和描述序列,包括发布日期信息、序列资料、源信息描述、参考文献信息和特征信息(如编码区、调节信号等)。递交者被允许在成为EMBL数据库最终出版格式前修改和查看所输入数据。Sequin是第二种递交格式,它是由NCBI建立的一种新的适合多平台(Mac/PC/Unix)使用的软件工具,它不限于递交到GenBank,也可发送到EMBL和DDBJ。Sequin程序可由NCBI网站直接下载和得到有关安装指南。电子邮件是第三种递交方式。递交者若要发送25种或更多的相关序列(如不同物种中同一基因序列),可直接通过电子邮件将序列发到datasubs@ebi.ac.uk。对于一些产生大量序列的研究组,可直接向EBI申请专门的发送账户,这样该用户可通过FTP或电子邮件进行传送数据和更新数据。

利用FTP服务器进行数据库查询和数据传输是进入EMBL数据库的两种基本方法。简单序列查找(simple sequence search)和SRS是用户最常用的检索数据库的工具。简单序列查找可根据登记号码检索到50个条目,而SRS可使用户检索所有数据库信息。

4.2.1.2 GenBank

GenBank始建于1988年,它是一个综合性的公共核苷酸和蛋白质序列数据库,并提供相关的文献目录和生物学注释。GenBank由位于美国NIH属下国立医学图书馆的NCBI进行维护。GenBank的数据来源主要是各研究人员独自递交,或由一些大的测序中心成批发送。这些成批数据的来源主要为进行EST、GSS和HTG的测序中心,也包括美国专利和商标事务所(USPYO)的专利序列数据。GenBank整合发送到EMBL和DDBJ的序列,每日进行交换,以保证这三大数据库对全部序列信息的维护。近10年来,GenBank的数据呈指数级增长(参见图4-1)。例如,以2003年6月的136发行版(release version)为例,GenBank包含325亿个核苷酸和2 560万个不同序列,代表着11.9万种不同的物种。

GenBank数据库结构分布方式是按传统方法分成不同的区组(目前已有18个),大致按分类学进行区分,如细菌(BCT)、病毒(VIR)、灵长类(PRI)和啮齿类(ROD)。最近,这种分组方法为满足特殊项目的需要,设立了EST、GSS和HTG等组别。

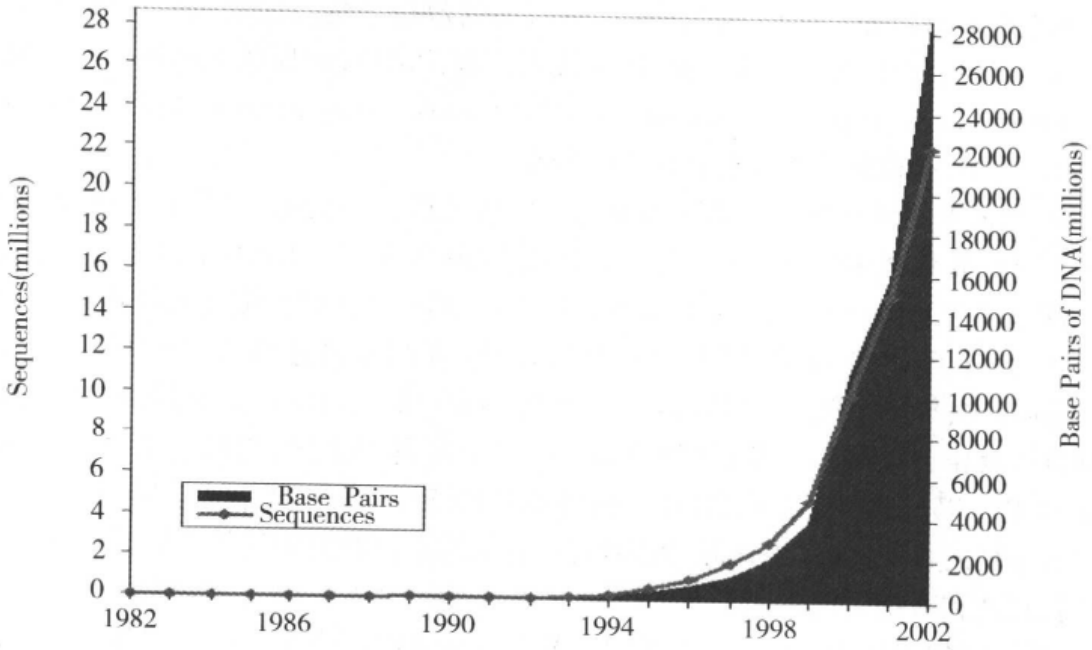


图 4-1　GenBank 数据库的增长曲线

资料来源:http://www.ncbi.nlm.nih.gov/

GenBank 的 flatfile(GBF)是 GenBank 数据库的基本单位。如附录 4-2 所示,GBF 可分为 3 个区:标题(header)、特征(feature)和序列(sequence)。标题区含有整个记录的描述信息。标题区的首行以 LOCUS 打头,包括其他成分,如登记号(accession number)(如 AF040078)、序列长度(如 1 932 bp)、分子类型(如 mRNA)、GenBank 部门(如 PLN),以及发布日期(如 16-JAN-1998)。GenBank 记录的定义行是用来概括该记录。该定义行出现于 FASTA 格式和 BLAST 检索结果中。第三行是登记号码(如 AF040078),该号码是专一的,用来代表该记录和数据库的主要参考代码。版本行由一系列标识符组成,包括登记号码和数据库中当前版本的版本号,接着是由 NCBI 指定每个序列的 GI 号码。关键词行是由一序列短语组成,概述基因产物和该记录的其他相关信息。标题区的其他行含有序列来源的物种名称或文献中的常用名称、作者、述评和备注等。特征区是表示该记录有关生物学信息的重要内容,如所编码的蛋白质或 RNA 分子的重要信息。最后的区是碱基的总数和各种碱基的数目以及首个碱基在基因组中的定位等信息。

事实上,所有进入 GenBank 的记录都是依靠直接递交进去,多数作者选用 Sequin 或 Bankit。Sequin 的方法在上面 EMBL 一节中已提及。Bankit 占单个递交的 65% 以上。使用 Bankit 时,作者根据表格内容逐项填写和修改并进行加注,特别适合单个或少量未完全注释好的递交,而且是 WWW 上使用的工具,不需要使用序列分析工具。

有 3 种常用的方法对 GenBank 数据库检索:Entrez[7]、BLAST[8,9] 序列相似性查找和 FTP 文件交换形式。Entrez 和 BLAST 的检索方法见 4.5 节。GenBank 数据库中的资料也可通过 FTP 方式从 NCBI 服务器(ftp://ncbi.nlm.nih.gov)上直接获得相关的 flatfile 格式。通过 FTP 既可得到全发行版形式(在'genbank'目录下的压缩文件),也可得到每日更新的文件(累积性更新文件在'daily'子目录下,非累积性更新文件在'daily-nc'值目录下)。对应于

BLAST 查找结果的数据集，要想获得只含序列的 FASTA 格式和非冗余性核苷酸和蛋白质序列，可在'BLAST/db'的子目录下获取。

4.2.1.3　DDBJ

DDBJ 代表的是日本 DNA 数据库。主要收集来自日本的数据，但也接受世界各地研究人员递交的数据。该数据库 1986 年开始使用，由日本教科文体部授权国立遗传研究所(NIG)建立。到 2003 年 6 月所发行的 DDJB 第 54 版中，DDBJ 已收集条目超过 2 500 万条，所含序列超过 320 亿个核苷酸，排位于前十名的物种分别是人、小鼠、大鼠、斑纹鱼、果蝇、水稻、玉米、鸡、拟南芥和芸苔。

DDBJ 文件分布与 GenBank 相似，划分为不同区组(division)，例外的是将"人类"从灵长类单列出来。DDBJ 的基本信息单位采用 GenBank 的 flatfile 格式。有 3 种不同的递交数据方法：第一种是通过 WWW 服务器利用 Sakura 递交；第二种是通过 E - mail 利用 MSS (mass submission systems)发送大量数据；第三种方法是按前面所述的 Sequin 方法。

进入 DDBJ 的方法有几种，"Getentry"是最常用的检索软件，可让用户按登记号码、位点名称、基因名称、产物名称、蛋白质序号、克隆名称和版本号等查找。其他方法可采用关键词的 SRS 检索法、表达序列的"SQ match"检索法、同源序列的 FASTA 和 BLAST 查找法，以及根据数据库分类的"TX Search"法等。

4.2.2　蛋白质序列数据库

蛋白质序列数据库(protein sequence database)以前只包括由实验获得的蛋白质或氨基酸序列，但目前所含的序列大部分是根据 DNA 序列获得的 ORF(多经人工校阅)。主要蛋白质序列数据库有：经多重比对获得的无空位的蛋白质高度保守区的短序列数据库 BLOCKS(http://www.blocks.fhcrc.org/)，G - 蛋白偶联受体数据库 GCRDB (http://receptor.mgh.harvard.edu/GCRDBHOME.html)，高质量校阅过的蛋白相互作用数据库 MIPS(http://www.mips.biochem.mpg.de/)，氨基酸突变数据库 PMD(http://pmd.ddbj.nig.ac.jp/)，用于蛋白质家族分析的高保守基序数据库 PRINTS (http://bioinf.man.ac.uk/dbbrowser/PRINTS/)，进行功能性注释的蛋白质序列数据库 PIR - PSD (http://pir.georgetown.edu/pir/)，以及高水平注释和校阅的蛋白质序列数据库 SWISS - PROT/TrEMBL(http://us.expasy.org/sprot/)。其中以 PIR - PSD 数据库和 SWISS - PROT 数据库最常用。

4.2.2.1　SWISS - PROT 和 TrEMBL

SWISS - PROT 是一种经校阅过的蛋白质序列数据库，首先于 1978 年在瑞士日内瓦大学医学生化系建立，随后与欧洲分子生物学实验室(EMBL)合作，目前由 EMBL 和新成立的瑞士生物信息学研究所(SIB)共同维护。2003 年 8 月，SWISS - PROT 的 41.21 发行版中，含有从超过 10 万条参考文献中得到的 133 312 个条目。图 4 - 2 显示该数据库的增长情况。

SWISS - PROT 数据库的重要特征是其高水平的注解、很低的冗余序列以及与其他数据库的高度整合。SWISS - PROT 的条目通常分为两类(见附录 4 - 3)：核心数据和注释。每个序列条目中的核心资料包括序列资料、参考文献、分类资料(描述蛋白质的生物来源)。注释资料则包括下列几项：①蛋白质功能；②翻译后修饰，如糖基化和磷酸化；③结构域和结合部位，如 ATP 结合部位；④次级结构，如 α - 螺旋、β - 片层；⑤四级结构，如同聚体或异聚

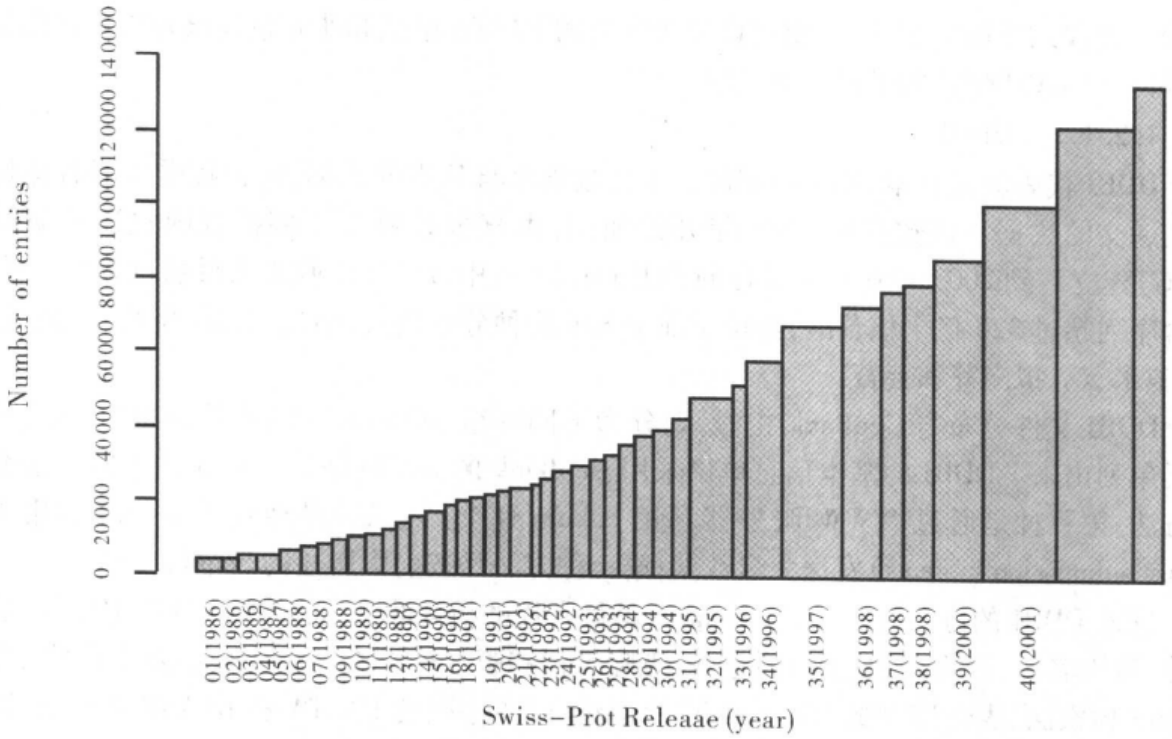


图 4-2　SWISS-PROT 数据增长曲线

资料来源:http://us.expasy.org/sprot/sprot-top.html

体;⑥与其他蛋白的相似性;⑦蛋白缺乏相关的疾病;⑧序列冲突或变异性。这些注释主要见于注解栏(CC)、特征表(FT)和关键词行(KW)。许多注解被分类为“主题(topic)”。条目中的每一行由 2 个字符代码组成,其代表的意义见表 4-1。

表 4-1　SWISS-PROT 数据库中的行代码

代码	意义	代码	意义
AC	登记号码	RX	参考文献相互参照
DT	日期	RA	参考文献作者
DE	描述	RT	参考文献题目
GN	基因名称	RL	参考文献位置
OS	物种名称	CC	注解或备注
OG	细胞器	DR	数据库相互参照
OC	物种分类	KW	关键词
OX	分类学相互参照	FT	特征表数据
RN	参考文献号	SQ	序列标题
RP	参考文献位置	(空白)	序列数据
RC	参考文献注解	//	结束行

为了加快处理基因组计划所带给 SWISS - PROT 的大量数据以及保证该数据库的高质量标准,已于 1995 年建立了一个 TrEMBL(translation from EMBL)数据库。TrEMBL 中的条目与 SWISS - PROT 的格式相同,其序列来自 EMBL 核苷酸数据库编码区的翻译序列,但不包括 SWISS - PROT 中已收录的序列。2003 年 8 月 15 日分布的 TrEMBL 发行版 24.8,含 939 599 个条目,其容量也在不断增加,新条目主要来自 EMBL 基因组计划的数据,且注释有限,如果经人工注释,则该条目从 TrEMBL 中移到 SWISS - PROT (图 4 - 3)。

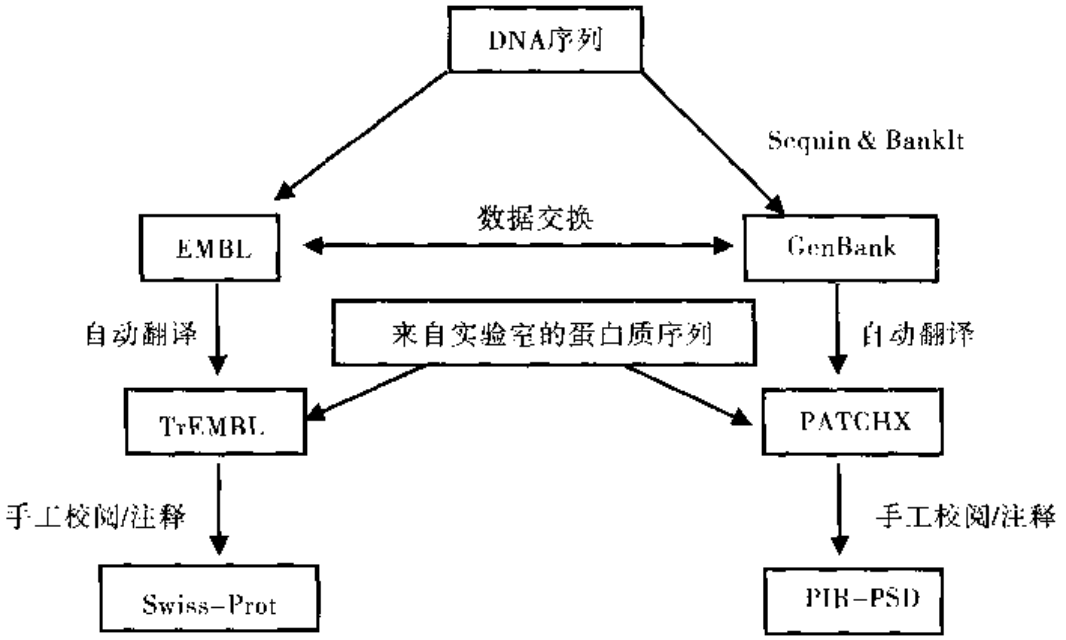


图 4 - 3 蛋白质序列数据库产生流程图

最有效的和最受用户青睐的交互式使用 SWISS - PROT 或 TrEMBL 数据库的方法是使用 ExPASy 网络服务器(http://www.expasy.org/)。该服务器提供文本检索方法,包括:①键入 SWISS - PROT 或 TrEMBL 中出现的任何文本;②通过描述基因名称和物种、作者、引文,以及表 4 - 1 中的代码行中的文本进行高级检索;③通过 SRS 软件进入数据库进行查询和检索所需序列条目。

4.2.2.2 PIR - PSD

PIR - PSD 是由 PIR 建立的最早的蛋白质序列数据库,其用途是帮助对蛋白质序列的鉴定和注释。它是由最初的 NBRF 蛋白质序列数据库[10]经过 20 多年的发展而由 Margaret O. Dayhoff演变为"蛋白质序列与结构图谱"(atlas of protein sequence and structure)[21]的。目前 PIR - PSD 由 NBRF、Munich 蛋白质序列信息中心(MIPS)和日本国际蛋白质信息数据库(JIPID)共同协作,收集已发表的蛋白质序列、来源、参考文献和特征信息等。PIR - PSD 是一种综合性的、非冗余的、组织得非常好又免费为用户使用的数据库,是对分子进化、功能基因组和计算生物学研究的有力支持。

PIR - PSD 数据库的增长速度比 SWISS - PROT 还要快,但每个条目只作较低水平的注释。该数据库由四部分组成:PIR_1 为完全归类好的条目;PIR_2 为已确认和归类的条目;PIR_3 是未确认条目;PIR_4 为非编码或非翻译条目。2003 年 7 月 14 日的第 77.02 发行版中,

PIR－PSD 包含 283 329 个记录，涵盖所有物种。数据库是属于非冗余的，完全相同或高度相似的序列被合并到同一条目。PIR－PSD 中的序列分类根据其相似性分成超家族、家族和同源结构域。其记录与其他主要的分子生物学数据库相互参照，如 Medline、GenBank、EMBL、DDBJ、PDB、HGDB 等。

与 SWISS－PROT 一样，PIR－PSD 是以其专家性注释而出名。PIR 的每个条目的注释内容包括许多原始递交记录中并不具有的一些重要特征，如图谱位置的遗传信息、内含子位置、起始密码子及与其他序列、结构、基因组和引文数据库的相互参照等。这些数据库包括许多公共的数据库，如 Entrez、Medline、PDB、GDB、OMIM、FlyBase、MIPS、SGD 和 TIGR 等。PIR－PSD 的另一重要特征是超家族的分类，以提供全面、非交叉性、等级聚类信息，并反映序列间的进化关系。已经人工校阅的超家族有数千个。到目前为止，PRI－PSD 序列的 99% 以上与相近序列被归类为同一家族（至少 45% 的序列是相同的），其中 2/3 的序列被归类到 36 000 个超家族。

进入 PIR－PSD 的方法有多种。第一种方法是根据文本查找记录和目录检索，该检索方法依照独特的标识符、登记号码和相互参照内容。第二种方法是 BLAST 检索，通过序列相似性检索数据库中的条目，一般使用 NCBI 的空位－BLAST 程序显示最高的 200 个匹配条目。第三种方法是“PIR pattern”检索。用户可自定义其类别，形成一个分子表达式来表示一组相关蛋白质的保守区，如[LIVM]－[VIC]－x(2)－G－[DENQTA]－x－[GAC]－x(2)－[LIVMFY](4)－x(2)－G，来检索数据库中是否具有该保守区序列。PIR 的无记名 FTP 站点（ftp://nbrfa.georgetown.edu/pir_databases）可提供 NBRF 和 CODATA 格式的 flat－file 文件转换，其相应序列为 FASTA/XML 格式或其相关的 DTD 文件格式。

PIR 也提供一种用计算机自动翻译的补充数据库 PATCHX，类似 SWISS－PROT 和 TrEMBL 的关系。新的条目主要是那些成批的来源于基因组测序计划的数据或选择性的 GenBank/EMBL 条目。PIR 的数据更新流程参见图 4－3。

4.3 基因组图谱数据库

关于遗传定位可追溯到上世纪初。Wilian Bateson 和 R. C. Punne 观察到甜豌豆的基因连锁现象。他们研究了两个基因：一种是花的颜色，另一种是花粉颗粒的形状。他们发现这两个基因存在物理关联性，但不知道这种关联的性质。其后，著名遗传学家摩根（Morgan）通过研究果蝇的两个常染色体基因证实了 Bateson 和 Punnett 的假说。其中一个基因决定眼睛的颜色，另一个基因决定翅长，F2 代中的表型数目严重背离了孟德尔规律所预计的数目。因此，摩根指出，决定这两个性状的基因位于同一对同源染色体上，并提出了著名的连锁与互换定律。

自从基因连锁规律发现以来，建立了遗传分析中不同的基因定位方法。这些定位方法不断向高分辨率方向发展。最基本的定位方法是将基因位点和分子标记分派（assign）到染色体的特定区域。通过计算两位点的遗传距离可得到高分辨遗传作图，进一步对 DNA 片段进行物理作图。

近年来，对于某个物种基因组的测序计划大大影响了遗传作图。许多情况下，研究人员只需要花费几小时检索一些定位数据库，就可锁定某个候选基因的关键区，而不需要像以前那样花费数月甚至几年的时间进行实验来确定。随着130多种微生物和其他物种数百个基因组全序列的完成，研究人员面临的任务是如何利用这些公共的基因组资源和相关的定位信息，有关定位克隆方法可参考3.5.3节。

4.3.1 基因组图的组成成分和类型

基因组图谱成分(elements)和分子标记(marker)显示DNA水平的多态性。因所采用的分子生物学方法不同，所得到的图谱成分和分子标记的类型也不同。从分子机制的观点来看，在DNA水平存在3种主要的多态类型：单碱基改变引起的单核苷酸多态性(SNP)，插入或缺失引起的片段长度多态性(FLP)，以及串联重复数目不同引起的多态性(VNTR)。若从单一位点来看，分子标记的类型也分为3种主要类型：双等位基因显性，如多态DNA随机扩增(RAPDs)和扩增片段长度多态性(AFLPs)；双等位基因共显现，如限制性片段长度多态性(RFLPs)；以及多等位基因共显性，如微卫星DNA。

指定基因或分子标记到某个染色体并确定其位置的方法有多种。经典的定位方法是配子定位(meiotic mapping)。其原理是根据双杂交或多杂交中染色体重组现象。例如根据自然发生的重组现象及其与已知位点的连锁关系来确定其排序。简单的方法是将一个携带未定位的等位基因的品系与携带分布于整个基因组、染色体位置已知的分子标记的一组品系进行杂交。通过观察多代中基因型来估计重组的可能性并进行位置估计(即计算遗传距离，参见第8章)。另一种配子定位的方法是利用RFLP标记和单序列长度多态性(SSLPs)标记。

第二类定位方法是通过原位杂交技术建立的细胞遗传学图谱(cytogenetic map)。定位于细胞遗传学图谱的标记可直接通过图像观察进行确定。常用的探针是通过放射性或荧光标记那些携带已知基因或分子标记的大片段插入的克隆。在荧光原位杂交(FISH)中，DNA探针经荧光标记，与部分变性的染色体制片温育。探针结合到该基因所在的染色体特定部位，其位置通过荧光斑点显现出来。

第三种类型的定位图为物理图(physical map)。这种图谱可以直接确定不同基因组成分的距离或根据克隆DNA片段对各基因组成分进行排序。最常采用的物理制图方法是利用序列标签位点(sequence－tagged site, STS)进行作图。

第四种类型也是最常用的图谱，是全基因组序列图(sequence map)。这种图谱将已知基因和基因间序列精确地排列到染色体上。

4.3.2 定位数据库

4.3.2.1 NCBI定位数据库

NCBI除了提供最著名的GenBank和Medline数据库外，也提供其他多种功能的数据库。其中一些专门为基因组定位而设计。这包括利用Entrez进入800多种物种的定位数据而进行的检索，即Entrez Map Viewer工具。它是一种交互式的物理图、遗传图、基因组序列、基因和其他基因组注解的查看工具，如非冗余性的STS数据集UniSTS、链接到不同数据库界面的LocusLink，以及人类基因图GeneMap'99等。

Entrez Map Viewer 工具根据染色体位置查找和显示基因组信息。对于感兴趣的区域可提供文本查询(如基因和分子标记名称)或根据序列比对(BLAST)查找。该查看工具可查看和检索某物种的整个基因组,显示出染色体图,可放大到某个细小区域,直到其序列信息。Entrez Map Viewer 在 4 个水平进行显示。第一个水平是在 NCBI 的主页,又称 NCBI Map View,可显示各种库存物种图谱,根据各物种目录链接到基因组图。第二个水平的图形显示是将整个基因组展现为根据染色体大小的表意图(ideogram),又称 Genome View,可对该基因组图进行全面检索,这些数据可包括一系列图谱,如序列图、细胞遗传图、遗传图、放射杂交图(RH)等。第三个水平是 Map View。它依次显示某染色体上图谱的各组成成分。进入 Map View 的方法有多种,包括:①点击 Genome View 图形页中某个染色体号;②点击基因组水平检索结果的图表中图谱成分名称;③点击基因组水平检索结果图表中的图形名称;④通过 LocusLink 检索结果点击进入(详见下文)。图 4 -4 显示人 Y 染色体的 Map View 图。Entrez Map Viewer 第四个水平的查看是特定染色体区段的序列。显示图形包括基因的编码区、RNA 以及对该序列特征的注解。有两种常用的方法进入 Sequence View:一种是通过基因组水平检索结果,通过点击染色体表意图下面的 SV;另一种是点击 Map View 显示的染色体区段结果,点击感兴趣部位的 SV(如图 4 -4 所示)。图 4 -5 显示图 4 -4 中对应的锌指蛋白 10 kb 的序列,可利用序列左、右两端的箭号,分别查看上、下游序列。有关 NCBI 基因组图谱其他特征的查找方法可参阅 3.6 节。

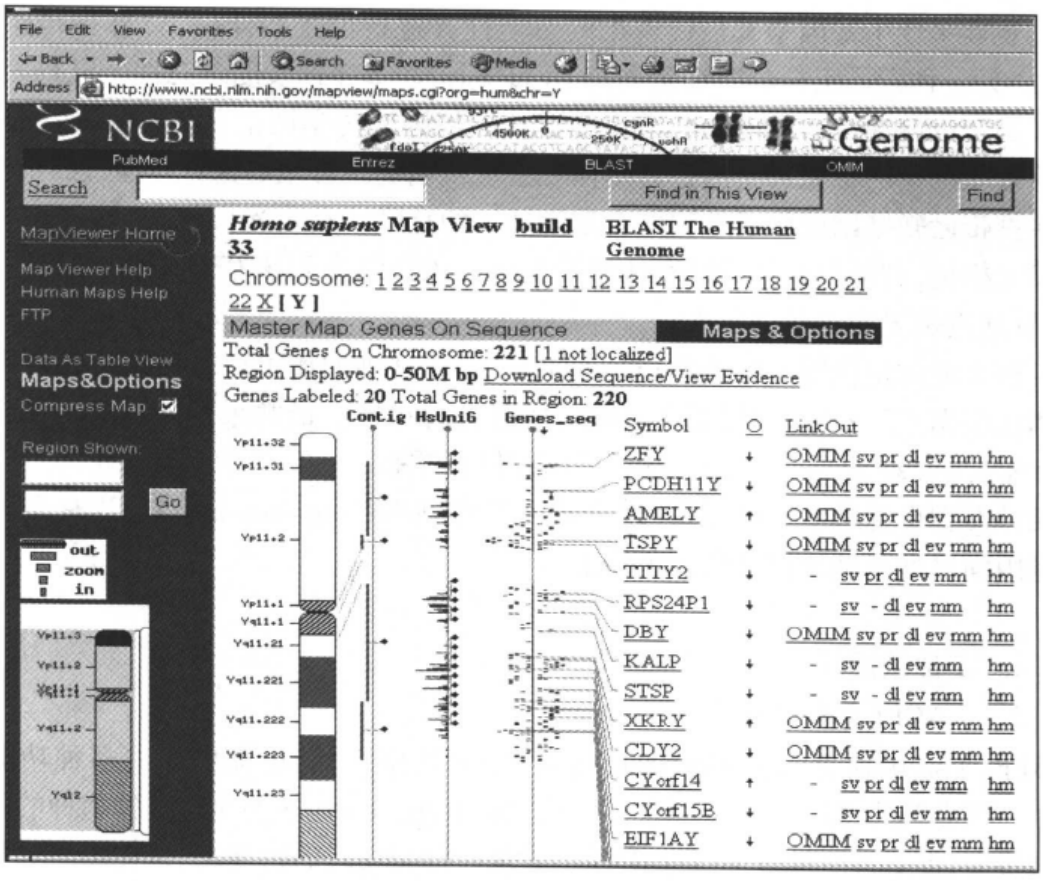


图 4 -4　通过 NCBI 的 Map View 得到的人 Y 染色体图

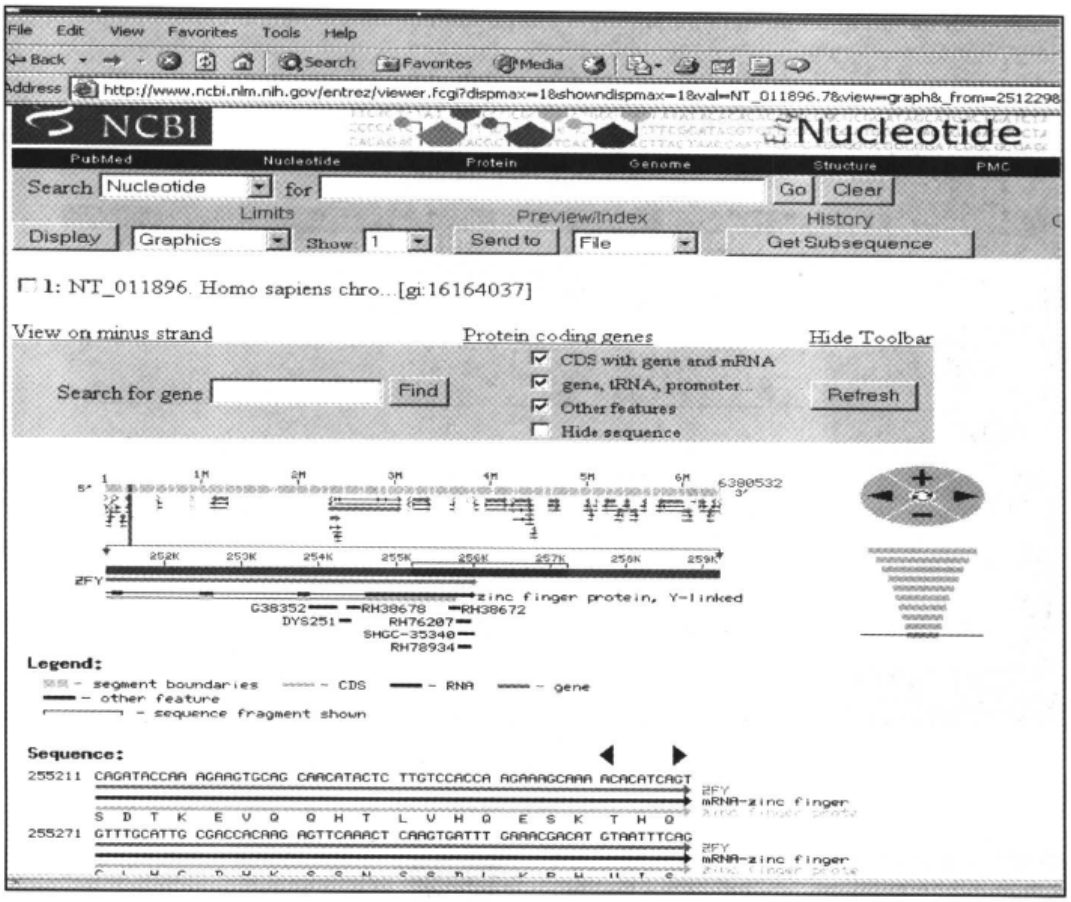


图4－5　点击图4－4中Y染色体图中锌指蛋白旁边的SV得到10 kb的Sequence View图

LocusLink（http://www.ncbi.nlm.nih.gov/LocusLink/）是另一种常用的检索工具，它可提供一个查询界面，分别链接到对指定位点的不同的信息类型，如表型、图谱位置、与其他基因的同源性等。LocusLink查询方式可以是文本（如蛋白质或疾病名称）、基因符号、序列登记号和数据库标志号（如MIM或EC号码）。LocusLink还支持其他更复杂类型的检索方法，如通用符号（wild card）（如ABC＊）、区域限制（如123456[mim]）和逻辑符号（如ABC＊和17[chr]）。LocusLink查找的第一步是在主页上方的查询框中键入文本术语，也可按字母顺序从基因列表中输入感兴趣的基因。以鸟苷酸结合蛋白－5（GNB5）为例，通过LocusLink得到图4－6的检索结果，分别显示人、小鼠和大鼠的相应基因条目。显示结果中第一列为LocusID号，是一个特定基因的标志号码，其他列依次显示物种、基因符号、位点描述、染色体定位和相关链接。这些链接中，由不同颜色的字母表示，P（深色）代表PubMed，O代表OMIM，R代表RefSeq，G代表GenBank，P（浅色）代表蛋白质序列，H代表同源序列，U代表UniGene，V代表SNP等不同数据库。

当点击LocusID号为14697时，显示小鼠的相关信息（图4－7）。结果上方用不同颜色的按钮链接外部数据库，如PubMed，UniGene，MAP，HOMOL，MGI等。报告结果分8个部分，分别提供有关LocusID和类型、功能、图谱信息、参考文献和相关序列、与其他数据库的链接等。

除上述Map View和LocusLink外，NCBI中还有其他与基因定位有关的工具，包括：

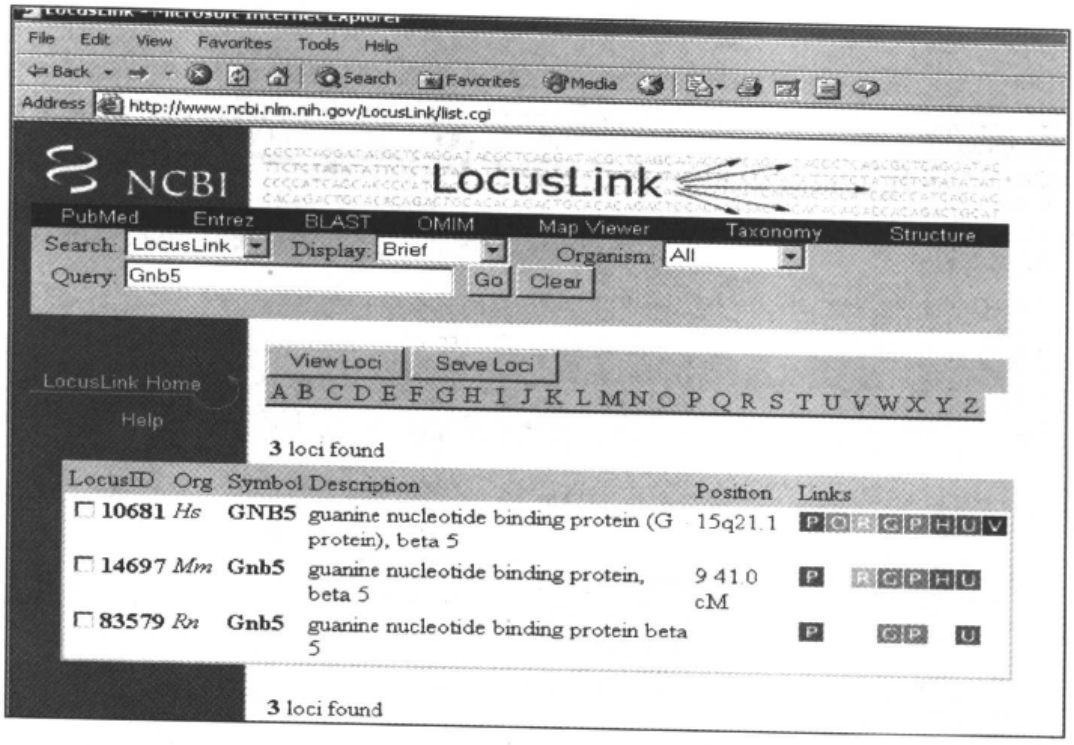


图 4 – 6　LocusLink 查询 GNB5 的结果显示图

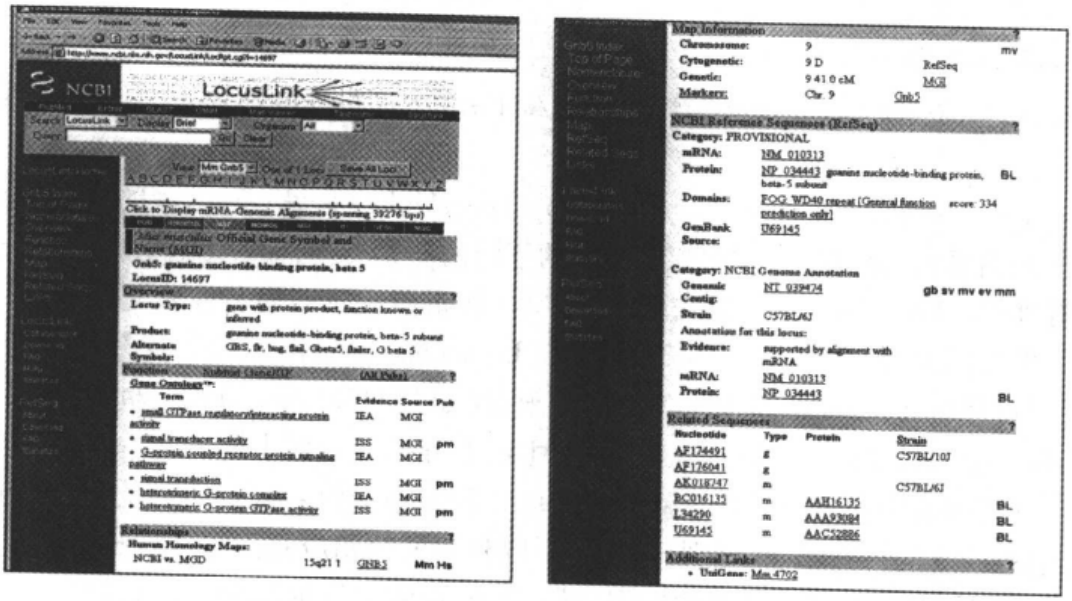


图 4 – 7　小鼠 GNB 5 的 LocusLink 报告结果

GeneMap99 http://www.ncbi.nlm.nih.gov/genemap99/

The Cancer Genome Anatomy Project http://www.ncbi.nlm.nih.gov/ncicgap/

Gene Nomenclature Resources http://www.ncbi.nlm.nih.gov/LocusLink/LLnomen.html

OMIM http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Entrez SNP http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp

dbSTS http://www.ncbi.nlm.nih.gov/dbSTS/

4.3.2.2　MGD

小鼠基因组数据库(mouse genome database, MGD)是一种综合性的数据库,由位于美国缅因州的 Jackson 实验室建立和维护,主要整合实验小鼠的遗传和基因组信息[11],包括小鼠基因、遗传标记、基因组特征、分子片段(探针、引物、cDNA 克隆、BAC 克隆、YAC 克隆等)、突变表型、比较基因组定位图、连锁图、细胞遗传图、物理图、实验定位资料(如 RH 定位)。自该数据库在 WWW 公布以来,其数据库内容不断更新和扩展,资料处理和检索工具也在不断改进。到 2003 年 8 月,该数据库包含 32 380 个基因,其中 30 270 个基因已被定位,29 177个基因具有 DNA 序列,5 602 个基因具有蛋白质序列,9 951 个基因具有人的同源基因,超过 29 130 个基因具有分子探针和片段资料,超过 3 600 个基因具有分子多态性。这些基因中,已鉴定 53 310 个分子标记,其中 49 850 个已定位,超过 12 740 个标记具有其多态性。

MGD 中的图谱和定位资料可通过不同方法获取,主要有"Mapping Data Query Form", "Recombinant Inbred Strain Distribution Patterns Query Form", "Recombinant Congenic Strain Distribution Patterns Query Form", "DNA Mapping Panel Data Sets", "Cytogentic Map"和"Physical Map"。另外,比较基因组图谱和数据可通过"Mammalian Orthology Query Form", "Whole Genome Orthology Map to Human or Rat"等方式获得,同时可得到染色体比较的"Oxford Grid"的数据矩阵。图 4-8 是一个"Mapping Data Query Form"检索 Ctla4 基因的例子。结果显示包括符号、名称、图谱位置、哺乳类同源性、序列、表型、基因分类、蛋白结构域、参考文献等概括性信息。许多详细信息可链接到外部数据库,如 GenBank, SWISS-PROT, TrEMBL, LocusLink, Ensembl, InterPro 等。MGD 同时提供小鼠与其他哺乳动物如人、大鼠等同源性比较资料和有关人与小鼠相同表型的注释。点击链接点可得到更详细的相关信息。此外,MGD 可按用户需要构建特殊区域的图谱,如整合个人的研究资料到连锁图等。

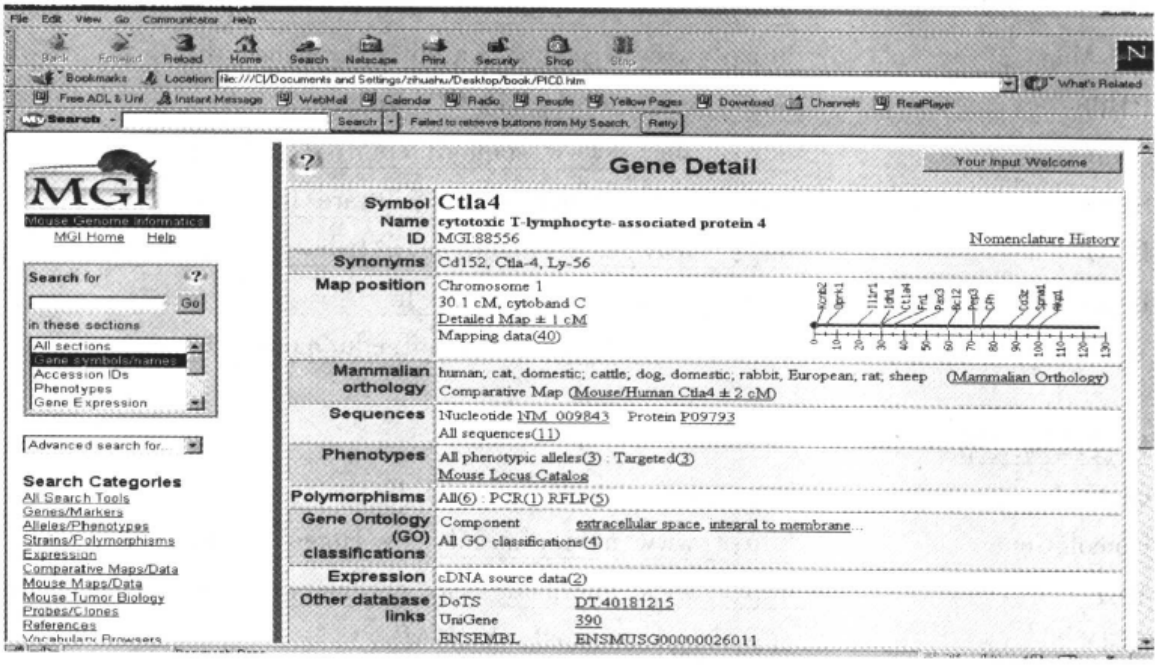


图 4-8　MGD 定位示意图

4.3.2.3 其他定位数据库

除上述大型定位数据库如 NCBI 和 MGD 外，其他一些定位数据库侧重于某些特殊方面，如有关遗传性和获得性染色体异常的细胞遗传学图谱、遗传连锁图、放射杂交图等。一些主要定位数据库和项目的网络资源如下：

遗传图

GDB	http://www.gdb.org/
RGD	http://rdg.mcw.edu
CHLC	http://gai.nci.nih.gov/CHLC/
Genethon	http://www.cephb.fr/ceph-genethon-map.html
HuGeMap	http://www.infobiogen.fr/services/Hugemap/
Marshfield Map	http://research._marshfieldclinic._org/genetics/Map_Markers/maps/IndexMapFrames.html
Map-O-MAT	http://compgen.rutgers.edu/mapomat/

物理图

CEPH	http://www.cephb.fr/ceph-genethon-map.html
HuGeMap	http://www.infobiogen.fr/services/Hugemap/
Whitehead map	http://www-genome.wi.mit.edu/cgi-bin/contig/phys_map
WICGR Human	http://www-genome.wi.mit.edu/cgi-bin/contig/phys_map
WICGR Mouse	http://www-genome.wi.mit.edu/cgi-bin/mouse/index

放射杂交图

GeneMap'99	http://www.ncbi.nlm.nih.gov/genemap99/
RH Map	http://www-genome.wi.mit.edu/mouse_rh/
RHdb	http://corba.ebi.ac.uk/RHdb/
Stanford	http://shgc-www.stanford.edu/RH/index.html
Sanger	http://www.sanger.ac.uk/Software/RHserver/
RHO	http://www.cs.technion.ac.il/Labs/cbl/CGI/rh-wizard.pl
ImpRH	http://imprh.toulouse.inra.fr/
WICGR	http://www-genome.wi.mit.edu/mouse_rh/

比较基因组图

CHLC	http://gai.nci.nih.gov/CHLC/
HomoloGene	http://www.ncbi.nlm.nih.gov/HomoloGene/
MGI	http://www.informatics.jax.org/menus/homology_menu.shtml
RGD	http://rgd.mcw.edu/VCMAP/
GDB	http://www.gdb.org/

整合图

GDB	http://www.gdb.org/
RGD	http://rgd.mcw.edu
UDB	http://genecards.weizmann.ac.il/geneloc/
LDB	http://cedar.genetics.soton.ac.uk/public_html/

细胞遗传图

CGAP	http://cgap.nci.nih.gov/
Coriell Cell Repositories	http://locus.umdnj.edu/
GDB	http://www.gdb.org/
NCBI	http://www.ncbi.nih.gov/genome/cyto/hbrc.shtml
NCBI	http://www.ncbi.nlm.nih.gov/sky/
WICGR	http://www.genome.wi.mit.edu/seq/mapping.html

植物遗传图

Rice(水稻)	http://www.tigr.org/tdb/e2k1/osa1/
Arabidopsis(拟南芥)	http://www.arabidopsis.org/servlets/mapper
Oryza sative (水稻)	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4530
Avena sativa (燕麦)	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4498
Glycine max (大豆)	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3847
Hordeum vulgare (大麦)	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4513
Triticum aestivum (小麦)	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4565
Zea mays (玉米)	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4577

4.4 医学数据库

尽管有关DNA和蛋白质序列的数据库是生物学家非常有用的工具，但是那些非序列数据的信息资源和有关基因突变的医学数据库也是研究基因和相关疾病的重要资源，这也是临床生物信息学(clinical bioinformatics)的重要内容之一。由于基因蕴含蛋白质的编码信息，蛋白质是体内生命活动的主要体现者。基因序列的变异影响着蛋白质的结构和功能，其

中一些可导致相应的表型即遗传病。本节主要讨论的常用医学数据库包括在线人类孟德尔遗传数据库 OMIM（http://www.ncbi.nlm.nih.gov/omim/），人类基因组变异数据库 HGVbase（http://hgvbase.cgb.ki.se/），人类基因突变数据库 HGMD（http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html）和 NCBI 中的 dbSNP 数据库（http://www.ncbi.nlm.nih.gov/SNP/）。

4.4.1 OMIM

OMIM（online Mendelian inheritance in man）是美国 Johns Hopkins 大学的著名遗传学家 Victor McKusick 建立的有关人类基因和遗传病目录的电子版[12]。它具有大量的可检索和随时更新的有关人类基因、遗传性状或遗传病的数据。OMIM 除提供相关的突变序列信息外，也收集来自公开发表的文献中具有遗传基础的人类疾病的信息及其临床特征和相关参考文献等。OMIM 多数有关基因的条目中，包含了已发表的等位基因或病理性基因突变的信息。截至 2003 年 9 月，OMIM 共有 14 730 个条目，其中 10 955 个为基因位点，1 387 个为表型的记录。

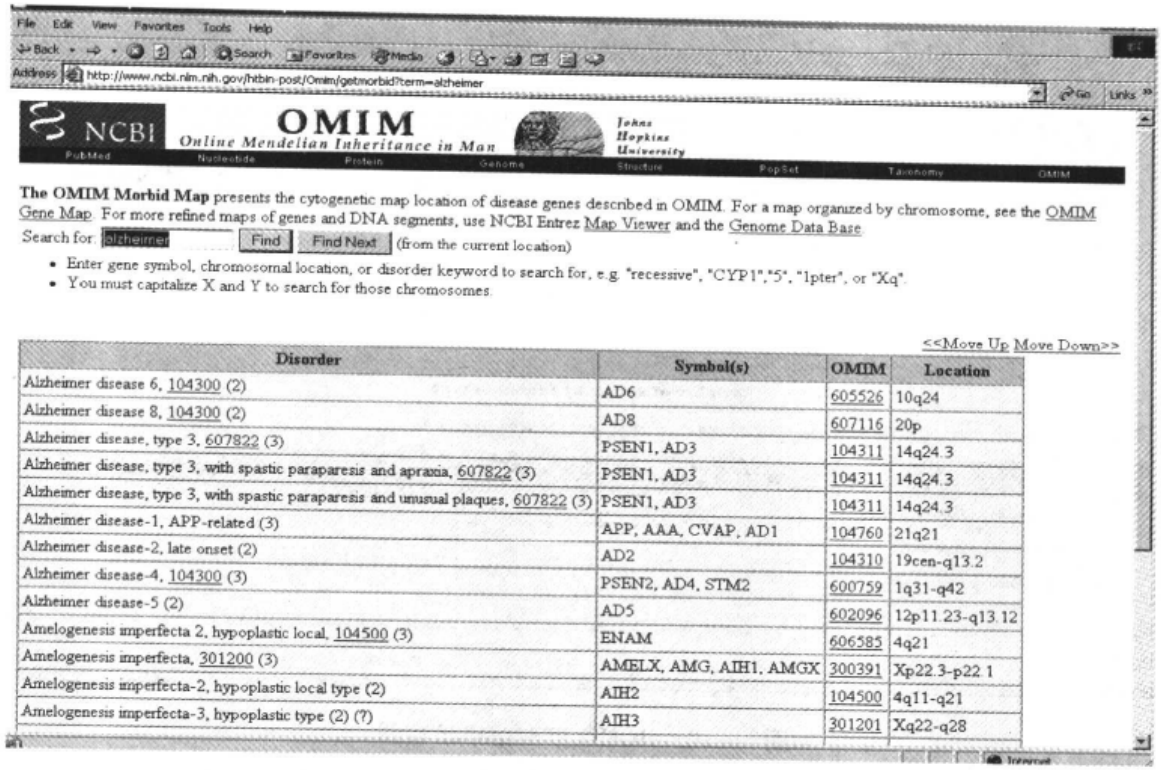
OMIM 也提供检索基因和遗传病的 3 个特殊的界面：OMIM 检索、OMIM 基因图和 OMIM Morbid 图。用户可在 OMIM 检索框中输入一个或多个关键词（包括临床、生化和细胞遗传特征等）。OMIM 基因图产生一个单一文档，显示该基因或位点的染色体和细胞遗传定位，有利于基因序列的观察。OMIM Morbid 图得到另一个单一文档，按字母顺序显示遗传病名和相应的细胞遗传图谱。

每个 OMIM 记录的显示结果中，包括 MIM 号、正式基因符号、基因产物名称、基因位点（如果已知的话）、疾病概述或基因产物功能，以及从发表文献中总结的有关基因、疾病或性状的文字信息等。OMIM 也链接到 Medline 的引文和其他相关的 OMIM 条目或 NCBI 的有关数据库。OMIM 基因图工具细胞遗传定位排列，分别显示基因符号、名称、MIM 号、相关疾病名称（如果已知）及其他信息。每个 OMIM 基因图条目都与 NCBI 的 Map View 和其他相关 OMIM 条目链接。OMIM 的 Morbid 图的记录中，按字母顺序将疾病名称排列，包括基因符号、疾病缩写、MIM 号和与相关基因和染色体位置的链接（参见图 4－9）。

4.4.2 HGMD

HGMD（human gene mutation database）是用来收集与人类遗传病相关的基因突变的数据库，由英国 Wales 大学医学院医学遗传研究所建立，目前与 Celera Genomics 合作，并得到 Genome Databases（GDB）等公司的支持。该数据库的初衷是用来研究人类基因突变的机制[13]，目前其用途更加广泛，包括不断更新的综合性人类基因病理改变的图谱，为基因诊断提供重要资源信息，以满足下列人员的需求：①人类分子遗传学研究人员和病理诊断学者；②对具有某种遗传病的患者或家庭感兴趣的医生；③遗传咨询师。

HMGD 含致病基因编码区的不同类型的突变，包括：编码区的单碱基置换、调节区和剪接区的突变、微小缺失或插入（indel）、三核苷酸重复扩增、大片段插入或缺失、重复和重排等。HGMD 中涉及的突变多引起明显的表型结果，还包括杂志中已发表与疾病相关的多态位点。这种疾病相关多态位点约占 HGMD 突变条目的 1.5%，多数为单核苷酸置换，少数为



Disorder	Symbol(s)	OMIM	Location
Alzheimer disease 6, 104300 (2)	AD6	605526	10q24
Alzheimer disease 8, 104300 (2)	AD8	607116	20p
Alzheimer disease, type 3, 607822 (3)	PSEN1, AD3	104311	14q24.3
Alzheimer disease, type 3, with spastic paraparesis and apraxia, 607822 (3)	PSEN1, AD3	104311	14q24.3
Alzheimer disease, type 3, with spastic paraparesis and unusual plaques, 607822 (3)	PSEN1, AD3	104311	14q24.3
Alzheimer disease-1, APP-related (3)	APP, AAA, CVAP, AD1	104760	21q21
Alzheimer disease-2, late onset (2)	AD2	104310	19cen-q13.2
Alzheimer disease-4, 104300 (3)	PSEN2, AD4, STM2	600759	1q31-q42
Alzheimer disease-5 (2)	AD5	602096	12p11.23-q13.12
Amelogenesis imperfecta 2, hypoplastic local, 104500 (3)	ENAM	606585	4q21
Amelogenesis imperfecta, 301200 (3)	AMELX, AMG, AIH1, AMGX	300391	Xp22.3-p22.1
Amelogenesis imperfecta-2, hypoplastic local type (2)	AIH2	104500	4q11-q21
Amelogenesis imperfecta-3, hypoplastic type (2) (?)	AIH3	301201	Xq22-q28

图4-9　OMIM 中有关 Alzheimer 氏病的基因定位列表

插入或缺失多态性。到2003年7月止，HGMD 包括1 399个基因中35 114个突变，并提供相应的1 243个 cDNA 序列。

HGMD 可通过查找界面按疾病名、基因名和基因符号进行查询。每个检索结果会得到对应的一系列基因符号。点击每个基因符号可进入一个记录，总结突变类型、表型、突变位置，以及与突变类型和表型有关的记录数目（参见图4-10）。点击突变类型将显示出登记号、位置、相关表型和与该突变相链接的参考文献。每个基因的记录还与突变图谱、该基因的 cDNA 序列、其他数据库中基因有关的记录相链接。

4.4.3　其他突变数据库

在许多其他基因突变资源中，HGVbase（human genome variation database）和 NCBI 的 dbSNP 最受研究人员的青睐。这两个数据库均收集常见的 SNP 数据。SNP 的发生频率为100～300个碱基中就有一个，这对遗传学中关联分析非常有用。

HGVbase 是由瑞典 Karolinska 基因组与生物信息学研究中心、EBI 和 EMBL 共同建立。该数据库的记录为人类基因组已知序列变异进行注释，这为研究人员了解基因组序列变异性如 SNP，及其与复杂性状的易感性或药物反应性，提供了重要的资源（分别参见3.1.3和9.2节）。每个 HGVbase 记录的资料来源于公开发表的基因组数据或文献，每个记录经人工校阅和注解。HGVbase 与 NCBI 的 dbSNP 共享数据，并整合 dbSNP 的一些记录。

dbSNP 是 NCBI 中建立的专门收集序列变异，如小片段插入或缺失、重复片段多态性和微卫星的公共数据库。其数据来源可以是由研究人员递交到数据库的任何物种的序列变异。该数据库从 GenBank 分离出来，但与 NCBI 的其他资源如 GenBank，LocusLink，PubMed

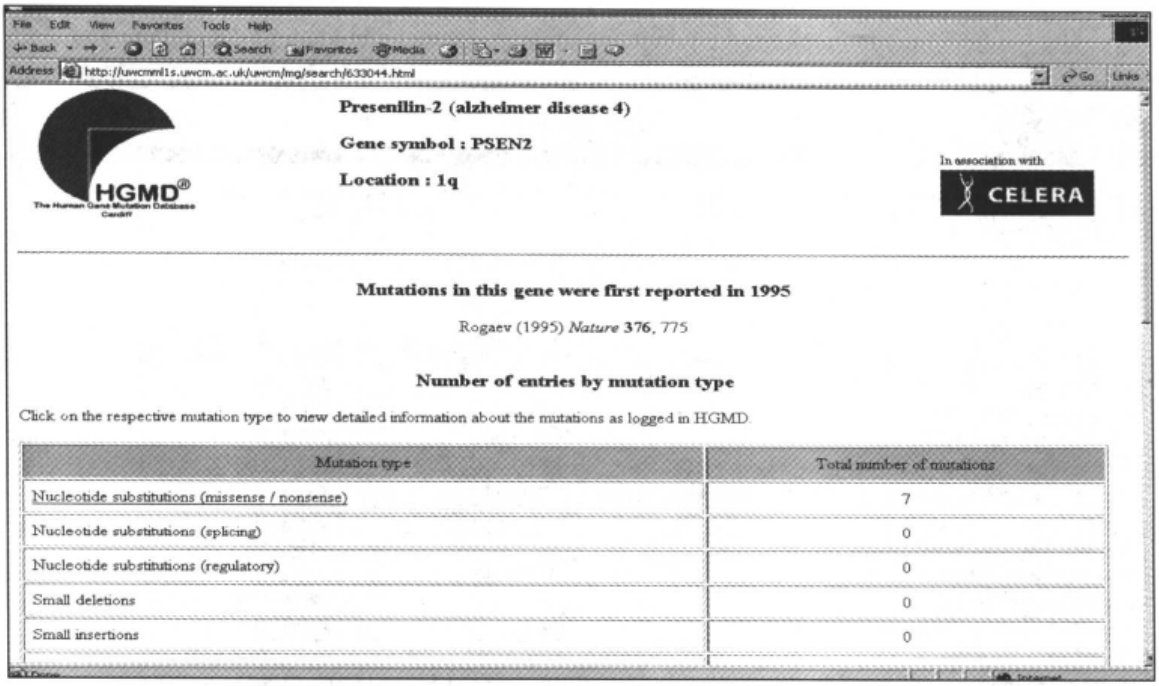


图 4－10　HGMD 查询结果示意图

图中显示与 Alzheimer 氏病有关的疾病、基因符号、报告日期、突变类型等结果

等相互链接。dbSNP 可通过 Entrez SNP 或 dbSNP 的“Easy Search Form”进行检索，也可通过 BLAST 检索工具来比较查询序列与数据库中已有序列的不同之处。

除上述数据库外，其他医学数据库列出如下：

遗传病发病率　http://archive.uwcm.ac.uk/uwcm/mg/fidd/

哮喘基因　http://cooke.gsf.de/wjst/home.cfm

乳腺癌基因　http://condor.bcm.tmc.edu/ermb/bcgd/bcgd.html

等位基因频率　http://alfred.med.yale.edu/alfred/index.asp

通用突变　http://www.umd.necker.fr/

蛋白质突变　http://www.genome.ad.jp/htbin/www_bfind?_pmd

疾病为中心的主要突变　http://www.genomic.unimelb.edu.au/mdi/dblist/disease.html

4.5　生物信息的获取与数据库查找工具

分子生物学数据库解决的是现代生物学中的关键性问题，即涉及 DNA 或蛋白质序列、功能基因组、蛋白组和生物大分子结构的相关信息。数据库的作用不是仅仅为了储存资料，而是需要获得和分发数据。GenBank 中的数据本身没有多大用途，除非能方便查找和获得一种可供使用而有意义的数据格式。已建立许多不同的算法和工具，从这些数据库中获取相关信息。这些方法大致分为两大类：第一类是根据文本的检索工具，如 Entrez；第二类是根据序列进行检索的工具，如 FASTA 和 BLAST。后一类主要根据序列相似性进行查找。这里主

要讨论上述3种分子生物学数据库的常用检索工具。有关基因组图的查找方法参见3.6节。定位数据库的查找，如Entrez Map View和LocusLink方法，已在4.3.2节中作过介绍。

4.5.1　Entrez

NCBI中的Entrez是最常用于获取分子生物学数据库信息的工具，如PubMed论文数据库、核苷酸和蛋白质数据库、三维结构数据库、基因组和定位数据库、基因表达数据库等（见图4－11）。Entrez提供一个综合数据库检索系统，可从超过13万个物种的DNA和蛋白质序列中获取相关信息。这些信息来源包括NCBI分类学、基因组图、人群、基因表达数据、分子模型数据库（MMDB）的蛋白结构、3D结构和结构域、PubMed、OMIM、在线书目等。PubMed包括1 200万个参考文献和从MEDLINE中得到的摘要，并与3 000多种可从网上获得全文的杂志建立链接。通过Entrez获得的序列资料来自GenBank，EMBL，DDBJ，RefSeq，PIR，PRF，SWISS—PROT，PDB等。

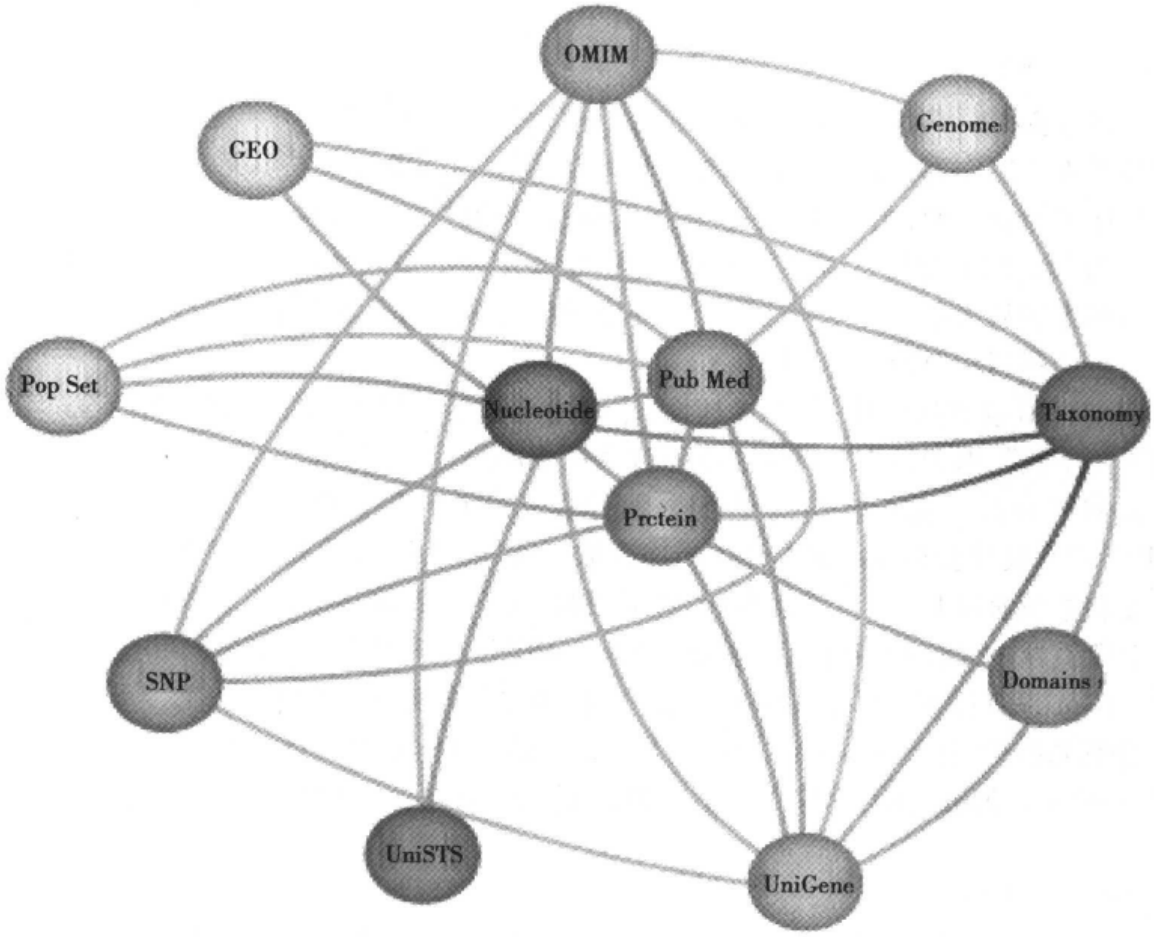


图4－11　Entrez综合信息检索系统示意图

数据来源：http://www.ncbi.nlm.nih.gov/Database/index.html

Entrez的查找可用不同的文本类型，如作者名、杂志名、基因或蛋白质名、物种、特殊标志（如GenBank登记号、序列ID、PubMed ID、MEDLINE UID），及根据所检索数据库不同而采用的其他形式。Entrez检索是根据不同公共数据库中不同条目之间，预先存在的逻辑相

互关系。因此,只要进行一个查询,可得到不同数据库之间存在相互联系的条目的综合信息。Entrez 的两种联系方式是近邻关系(neighboring)和硬链接(hard link)。近邻关系是预先运算的相似性查找方式,发现数据库记录中相关的记录。根据数据库的不同,其算法有所不同。例如,若用户想查找某个 PubMed 记录,用户可要求 Entrez 找出 PubMed 中与原文主题相似的其他文献。同样,若用户想要查询某序列,Entrez 可回复与该序列相似的一系列其他序列。近邻关系的确定是根据统计学方法测定其相似程度(参见 4.5.3"BLAST")。硬链接是用来将不同数据库之间存在逻辑关系的条目建立相互联系,使得不同数据库的资料得以整合。例如,一条 MEDLINE 建立中引用了 GenBank 中的核苷酸序列的记录,后者被链接到蛋白质序列,由此,3 个数据库之间就建立起某种联系。

通过 Entrez 获取的结果可以不同的格式显示或下载。格式的选用依记录类型而异。例如 GenBank 的记录格式分为 Flatfile, FASTA, XML, ASN.1 等。某些记录如基因组记录还提供图形格式。

4.5.2 FASTA

FASTA 是最早广泛使用的一种数据库相似性检索程序[14,15],用于检索核苷酸和蛋白质序列数据库中的序列相似性和同源性。FASTA 是专门为鉴定那些从远古祖先继承下来的蛋白质序列而设计的,尤其对那些演化程度高、低相似性的长片段更为适用。许多情况下,FASTA 查找的结果为与查询序列同源的具有很高记分的系列文库序列,或者得到相似记分的一系列无法从大量文库序列区分开来的序列。这两种情况下都能明白地回答文库中是否具有与查询序列相似的序列问题。

FASTA 检索方法快捷,具有选择性,主要由于其最初仅考虑氨基酸的相同性。其敏感性的提高不仅仅因为利用 PAM250 矩阵对大量相同氨基酸进行记分和重记分,同时也进行开始区域的合并。现行常用的 FASTA 程序是 FASTA3, FASTX3, TFASTA3 和 TFASTX3。FASTA3 用来比较蛋白质序列与蛋白质序列数据库,或者 DNA 序列与 DNA 序列数据库。检索速度和选项通过 ktup(文字大小)产生来控制:蛋白质比较用 1 和 2,DNA 比较用 1 到 6。选用的数值越小,敏感性越高,但速度越慢。FASTX3 用来比较 DNA 序列和蛋白质序列数据库。DNA 序列采用 3 种不同读框,允许空位和位移的存在。TFASTX 比较蛋白质序列与 DNA 序列数据库,计算正向和反向移码(frame shift)的相似性。TFASTA3 用来比较蛋白质序列与 DNA 序列数据库,计算 3 种正向和 3 种反向读框的相似性(无移码)。

4.5.3 BLAST

BLAST 是一种快速序列比较工具,采用启发式方法根据优化的局部相似性构建比对关系[8,9]。局部比对只比较两序列某些区域的相似性,而整体比对是比较两序列的全长序列(参见第 5 章)。BLAST 利用局部比对找出两序列的相似性区域。由于 BLAST 在比较 DNA 或蛋白质序列时较动态编程方法 Smith - Waterman [16] 和 Needleman - Wunsch [17] 快,因而广泛用于数据库的查找。

NCBI 的 BLAST 工具是最常用的站点。该程序可对一系列序列数据库进行序列相似性分析,查询方法可用序列或 GenBank 登记号。表 4 - 2 所列为不同的 BLAST 亚类,主要基于

所查询序列和检索的数据库不同而设计。

表4-2 BLAST检索程序的比较

程序	查询内容	数据库	比较方法	评价
BLASTP	蛋白质	蛋白质	蛋白质序列和蛋白质数据库的比较	利用置换矩阵比找出相互关系距离
BLASTN	核苷酸	核苷酸	DNA序列与DNA序列数据库的比较	
BLASTX	核苷酸（翻译的）	蛋白质	不同翻译阅读方式的核苷酸序列与蛋白质数据库的比较	方便找出未知核苷酸序列的可能蛋白质序列
TBLASTN	蛋白质	核苷酸（翻译的）	查询蛋白质序列与各种阅读方式的核苷酸序列数据库的比较	可用于数据库中未注释的编码序列
TBLASTX	核苷酸（翻译的）	核苷酸（翻译的）	查询核苷酸序列的6种阅读方式与核苷酸数据库的6种阅读方式的比较	常用于EST分析

BLAST检索的结果显示是查询序列与相似序列之间的可允许空位比对关系，其显示内容一般分为4个部分：最上方为独特的申请ID号（RID）、查询信息、数据库信息及其与分类学的链接；第二个部分为查询序列比对的图形显示窗；接下来的第三部分为找出的相似序列的描述；第四部分是两两比对的具体结果。

图4-12显示的是查询的luciferin 4-monooxygenase（P08659）第三、四部分的一些结果。第三部分结果显示栏的左侧为该序列记录的链接点。右侧的记分值是根据空位和置换得到的计算结果（参见第5章）。记分越高，比对的相似性越显著。每个记分链接到第四部分的两两比对结果。E值（expected value）表示具有相似记分的序列在数据库中出现的随机出现的可能性。E值越小，表明比对关系越显著。例如图4-12（a）中第一个比对的E值为0.0，表明具有同样记分的序列随机性是不可能的。第三部分最右边所显示的字符，是与对应数据库链接的标志，如L表示LocusLink，S表示NCBI中的MMD结构数据库。图4-12（b）表示的是BLAST结果的第四部分的内容，显示查询序列（Query）与所配对的序列（Sbjct）的比对，中间行显示相同或相似氨基酸。

(a)

Sequences producing significant alignments:	Score (bits)	E Value
gi|209229|gb|AAA72988.1| luciferase/kanamycin resistance pr…	1098	0.0
gi|13160953|gb|AAK13426.1|AF320510_4 luciferase [Promoter p…	1093	0.0
gi|126501|sp|P08659|LUCI_PHOPY Luciferin 4-monooxygenase (L…	1092	0.0
gi|17530182|gb|AAL40737.1| tissue factor/luciferase fusion…	1092	0.0
gi|1469266|emb|CAA59281.1| firefly luciferase [Photinus pyr…	1091	0.0
gi|3123921|gb|AAC40214.1| firefly luciferase [Reporter vect…	1090	0.0
gi|17530179|gb|AAL40735.1| protein serine kinase/luciferase…	1090	0.0
gi|348485|pir|S29354 Photinus-luciferin 4-monooxygenase (A…	519	e-146
gi|2133576|pir|S29355 Photinus-luciferin 4-monooxygenase (…	517	e-145
gi|478790|pir|S29353 Photinus-luciferin 4-monooxygenase (A…	516	e-145
gi|3929284|gb|AAC79854.1| luciferase [MMTV-luciferase repor…	463	e-129
gi|31201475|ref|XP_309685.1| ENSANGP00000003832 [Anopheles…	385	e-105
gi|31205637|ref|XP_311770.1| ENSANGP00000014318 [Anopheles…	380	e-104
gi|21355181|ref|NP_651221.1| CG6178-PA [Drosophila melanoga…	377	e-103
gi|31219007|ref|XP_316739.1| ENSANGP00000016100 [Anopheles…	371	e-101
gi|12229631|sp|O24145|4CL1_TOBAC 4-coumarate--CoA ligase 1…	292	1e-77

(b)

>gi|478790|pir|S29353 Photinus-luciferin 4-monooxygenase (ATP-hydrolysing) (EC 1.13.12.7)
[similarity]-luminescent click beetle (Pyrophorus plagiophthalmus)
Length=543

Score=516 bits(1329), Expect=e-145
Identities=253/545(46%), Positives=362/545(66%), Gaps=5/545(0%)

```
Query: 1    MEDAKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDAHIEVNITYAEYFEMS   60
            M+  KN+  GP P +PLED TAGE L +A++++ +P   A  D    + +++Y E+FE +
Sbjct: 2    MKREKNVIYGPEPLHPLEDLTAGEMLFRALRKHSHLPQ--ALVDVFGDESLSYKEFFEAT   59

Query: 61   VRLAEAMKRYGLNTNHRIVVCSENSLQFFMPVLGALFIGVAVAPANDIYNERELLNSMNI   120
              LA+++   G  N  + +C+EN+ +FF+P++ A +IG+ VAP N+ Y   EL  M +
Sbjct: 60   CLLAQSLHNCGYKMNDVVSICAENNKRFFIPIIAAWYIGMIVAPVNESYIPDELCKVMGI   119
```

图 4-12 BLASTP 检索结果报告格式

luciferin 4-monooxygenase(P08659)被用做 BLASTP 的查询 nr 数据库的条目。上图(a)显示找到的清单；下图(b)显示与 photinus-luciferin 4-monooxygenase 的比对

(胡自华 著 张 庆 李 巍 译)

参 考 文 献

1 Dayhoff MO, Eck RV, Chang MA, *et al*. Atlas of Protein Sequence and Structure Vol. 1. National Biomedical Research Foundation, Silver Spring, 1965

2 Fraser CM, Gocayne JD, White O, *et al*. The minimal gene complement of Mycoplasma genitalium. Science, 1995,270: 397

3 Flanders DJ, Weng S, Petel FX, *et al*. AtDB, the Arabidopsis thaliana database, and graphical – web – display of progress by the Arabidopsis Genome Initiative. Nucleic Acids Res, 1998,26: 80

4 Garcia – Hernandez M, Berardini TZ. TAIR: a resource for integrated Arabidopsis data. Funct Integr Genomics, 2002,2: 239

5 Stoesser G, Baker W, van den Broek A, *et al*. The EMBL nuceotide sequence database. Nucleic Acids Res, 2002,30: 21

6 Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. Methods Enzymol, 1996,266:114

7 Schuler GD, Epstein JA, Ohkawa H, *et al*. Entrez: molecular biology database and retrieval system. Methods Enzymol,1996,266: 141

8 Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. J Mol Biol, 1990, 215: 403

9 Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI – BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997,25: 3 389

10 George DG, Barker WC, Hunt LT. The protein identification resource (PIR). Nucleic Acids Res, 1986,14: 11

11 Blake JA, Eppig JT, Richardson JE, *et al*. The Mouse Genome Database (MGD): the model organism database for the laboratory mouse. Nucleic Acids Res, 2002,30: 113

12 McKusick VA. Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders. Baltimore: The Johns Hopkins University Press, 1998

13 Cooper DN, Krawczak M. Human Gene Mutation. Oxford: BIOS Scientific Publishers, 1993

14 Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science, 1985, 227: 1 435

15 Pearson W. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol, 2000,132: 185

16 Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol, 1981,147: 195

17 Needleman SB, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 1970,48: 443

附录4－1　EMBL Flatfile 格式

```
ID  HSIGHAF  standard; RNA; HUM; 1089 BP.
XX
AC  J00231;
XX
SV  J00231.1
XX
DT  13-JUN-1985 (Rel. 06, Created)
DT  04-MAR-2000 (Rel. 63, Last updated, Version 8)
XX
DE  Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW  C-region; gamma heavy chain disease protein;
KW  gamma3 heavy chain disease protein; heavy chain disease; hinge exon;
KW  immunoglobulin gamma-chain; immunoglobulin heavy chain;
KW  secreted immunoglobulin; V-region.
XX
OS  Homo sapiens (human)
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC  Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN  [1]
RP  1-1089
RX  MEDLINE; 82247835.
RX  PUBMED; 6808505.
RA  Alexander A., Steinmetz M., Barritault D., Frangione B., Franklin E. C.,
RA  Hood L., Buxbaum J. N.;
RT  "gamma Heavy chain disease in man: cDNA sequence supports partial gene
RT  deletion model";
RL  Proc. Natl. Acad. Sci. U.S.A. 79(10):3260-3264(1982).
XX
DR  ENSEMBL; ENSG00000130076; ENST00000251006.
DR  GDB; 119339; IGHG3.
DR  GOA; P01860.
DR  IMGT/LIGM; J00231; J00231.
```

```
DR  SWISS-PROT; P01860; GC3_HUMAN.
XX
CC  The protein isolated from patient OMM is a gamma heavy chain
CC  disease (HCD) protein. It has a large 5' internal deletion
CC  consisting of most of the variable region and the entire ch1
CC  domain. [1] suggests that the protein abnormality is from a partial
CC  gene deletion rather than from defective splicing.
XX
FH  Key Location/Qualifiers
FH
FT  source          1..1089
FT                  /db_xref="taxon:9606"
FT                  /mol_type="mRNA"
FT                  /organism="Homo sapiens"
FT                  /map="14q32.33"
FT  mRNA            <1..1089
FT                  /note="gamma3 mRNA"
FT  CDS             23..964
FT                  /codon_start=1
FT                  /db_xref="GOA:P01860"
FT                  /db_xref="SWISS-PROT:P01860"
FT                  /note="OMM protein (Ig gamma3) heavy chain"
FT                  /gene="IGHG3"
FT                  /protein_id="AAA52805.1"
FT                  /translation="MKXLWFFLLLVAAPRWVLSQVHLQESGPGLGKPPELKTPLGDTTH
FT                  TCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPRCPEPKSCDTPPPCPXCPAPELLGGP
FT                  SVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPXVQFKWYVDGVEVHNAKTKLREEQYN
FT                  STFRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPXXXXXXXXXXXXE
FT                  EMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYNTTPPMLDSDGSFFLYSKLTVDKS
FT                  RWQQGNIFSCSVMHEALHNRYTQKSLSLSPGK"
FT  sig_peptide     26..79
FT                  /note="OMM protein signal peptide"
FT                  /gene="IGHG3"
FT  mat_peptide     80..961
FT                  /note="OMM protein mature peptide"
FT                  /gene="IGHG3"
XX
SQ  Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
```

cctggacctc	ctgtgcaaga	acatgaaaca	nctgtggttc	ttccttctcc	tggtggcagc	60
tcccagatgg	gtcctgtccc	aggtgcacct	gcaggagtcg	ggcccaggac	tggggaagcc	120
tccagagctc	aaaaccccac	ttggtgacac	aactcacaca	tgcccacggt	gcccagagcc	180
caaatcttgt	gacacacctc	ccccgtgccc	acggtgccca	gagcccaaat	cttgtgacac	240
acctccccca	tgcccacggt	gcccagagcc	caaatcttgt	gacacacctc	ccccgtgccc	300
nnngtgccca	gcacctgaac	tcttgggagg	accgtcagtc	ttcctcttcc	ccccaaaacc	360
caaggatacc	cttatgattt	cccggacccc	tgaggtcacg	tgcgtggtgg	tggacgtgag	420
ccacgaagac	ccnnnngtcc	agttcaagtg	gtacgtggac	ggcgtggagg	tgcataatgc	480
caagacaaag	ctgcgggagg	agcagtacaa	cagcacgttc	cgtgtggtca	gcgtcctcac	540
cgtcctgcac	caggactggc	tgaacggcaa	ggagtacaag	tgcaaggtct	ccaacaaagc	600
cctcccagcc	cccatcgaga	aaaccatctc	caaagccaaa	ggacagcccn	nnnnnnnnnn	660
nnnnnnnnnn	nnnnnnnnnn	nnnnngagga	gatgaccaag	aaccaagtca	gcctgacctg	720
cctggtcaaa	ggcttctacc	ccagcgacat	cgccgtggag	tgggagagca	atgggcagcc	780
ggagaacaac	tacaacacca	cgcctcccat	gctggactcc	gacggctcct	tcttcctcta	840
cagcaagctc	accgtggaca	agagcaggtg	gcagcagggg	aacatcttct	catgctccgt	900
gatgcatgag	gctctgcaca	accgctacac	gcagaagagc	ctctccctgt	ctccgggtaa	960
atgagtgcca	tggccggcaa	gccccccgct	cccgggctct	cggggtcgcg	cgaggatgct	1020
tggcacgtac	cccgtgtaca	tacttcccag	gcacccagca	tggaaataaa	gcacccagcg	1080
ctgccctgg						1089

附录4－2　GenBank Flatfile 格式

LOCUS　AF040078　1932 bp　mRNA　linear　PLN 16－JAN－1998

DEFINITION　Prunus serotina (R)－(+)－mandelonitrile lyase isoform MDL2 precursor, mRNA, complete cds.

ACCESSION　AF040078

VERSION　AF040078.1 GI:2773273

KEYWORDS

SOURCE　Prunus serotina (black cherry)

　ORGANISM　Prunus serotina

　Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; rosids; eurosids I; Rosales; Rosaceae; Amygdaloideae; Prunus.

REFERENCE　1 (bases 1 to 1932)

　AUTHORS　Hu, Z. and Poulton, J. E.

　TITLE　Prunus serotina mRNA for (R)－(+)－mandelonitrile lyase isoform MDL2, cds

JOURNAL Unpublished
REFERENCE 2 (bases 1 to 1932)
AUTHORS Hu, Z. and Poulton, J. E.
TITLE Direct Submission
JOURNAL Submitted (24-DEC-1997) Biological Sciences, The University of Iowa, Iowa City, IA 52242, USA
FEATURES Location/Qualifiers
source 1..1932
/organism="Prunus serotina"
/mol_type="mRNA"
/db_xref="taxon:23207"
CDS 18..1748
/codon_start=1
/product="(R)-(+)-mandelonitrile lyase isoform MDL2 precursor"
/protein_id=""
/db_xref="GI:2773274"
/translation="MVKSTMSAILVLALHLFVLHLQYSEVQSLANTSAHDFSYLEFVY
DANDTELEGTYDYIIVGGGTAGCPLAATLSANYSVLVLERGTLPTEYPNLLTSDGFTY
NLQQEDDGQTPVERFVSGDGIDNVRGRVLGGTSMINAGVYVRANTSFFNQTGIEWDMD
LVNKTYDWVEDTIVFKPDFQFWQNLTGTAFLEVGILPDNGFSLDHLEGTRLTGSTFDN
NGTRHASDELLNKGDPNNLRVAVHAAVEKIIFSSDSSGVTAIGVIYTDSNGTTHQAFV
RGDGEVILSAGPIGSPQLLLLSGVGLESYLTSLNISVVASHPYVGQYTYDNPRNFINI
LPPNPIEASTVTVLGITSDFYQCSISSLPFSTAPFGFFPNPTYPLPNTTFAHIVNKVP
GPLSHGTVLLQSTSDVRVAPNVTFNYYSNTTDLAHCVSGMKKIGEFLSSDALKPYKVE
DLPGIEGFDILGIPLPENQTDDAAFETFCREAVASYWHYHGGCLVGEVLDDDFRVTGI
NALRVVDGSTFPSTPASHPQGFYLMLGRYMGTKILQERLASEEALHKSTFEPKILESL
ESALSFAFES"
sig peptide 18..101
mat peptide 102..1745
/product="(R)-(+)-mandelonitrile lyase isoform MDL2"
BASE COUNT 541 a 411 c 419 g 561 t
ORIGIN

1	gcacgagatt	cagaaacatg	gtgaaatcaa	caatgtcagc	tatactagta	ttggcgctgc
61	accttttgt	ccttcatctt	caatactcag	aggttcaatc	gcttgccaat	acctctgctc
121	atgattttag	ctacttggaa	tttgtttacg	atgccaatga	cacagagttg	gaaggaacat
181	acgactatat	tattgttggt	ggaggaacag	cagggtgtcc	attggcagct	actttgtcag
241	caaactactc	ggtgcttgtt	ctggaaaggg	gcactcttcc	tacagaatat	ccaaacttgt

301	tgacttcaga	tgggtttata	tataatctgc	agcaagaaga	tgatggacag	acaccagtcg
361	aaaggttcgt	gtctggagat	ggtattgaca	atgtaagagg	gagagtcctc	ggtggcacga
421	gcatgatcaa	tgccggggtc	tacgtaagag	ctaacacctc	gttctttaat	caaacaggga
481	ttgaatggga	catggatttg	gttaataaga	catatgactg	ggttgaagac	actattgtgt
541	tcaagccgga	tttccaattt	tggcaaaatc	ttacaggaac	tgcattcttg	gaggtaggta
601	ttcttccaga	caacggattt	agtttggatc	acctagaagg	aactagactc	accggctcaa
661	ctttcgacaa	taacggaacc	agacatgcat	ctgatgaact	gcttaataaa	ggagacccaa
721	acaacttgcg	agttgcagtt	catgccgcag	tagagaagat	catcttctct	tccgattcat
781	caggtgtgac	agctatagga	gttatatata	ctgattcgaa	cggaacgact	catcaagcat
841	ttgtacgcgg	tgacggagaa	gttatattga	gtgcagggcc	aattgggtcc	cctcaacttc
901	tactacttag	tggcgttggc	ctggagtctt	acctaacatc	actgaacatt	tcagttgttg
961	cttcccatcc	ttacgtcggg	cagtatatat	atgacaatcc	tcgtaatttc	attaacattt
1021	tgccccccaaa	tccaatagaa	gcctcaactg	tgactgttct	aggcattaca	agcgacttct
1081	accaatgttc	aatatcgagc	ttgccatttt	ctactgcacc	cttcggtttt	tttcctaatc
1141	caacctatcc	cctgccaaat	acaactttcg	ctcacattgt	taacaaagtt	ccgggaccac
1201	tatctcatgg	tactgtcctg	ctgcaatcaa	cctctgatgt	gagagtcgct	ccaaatgtca
1261	cattcaacta	ctactcgaat	acaacggacc	ttgctcattg	tgttagcggc	atgaagaaga
1321	ttggtgaatt	cttgagctca	gacgcattaa	aaccatataa	agtggaagat	ttgccaggta
1381	tagaaggttt	tgatatttta	ggaatacctt	tgccagagaa	ccagacagat	gatgcagcct
1441	tcgaaacatt	ttgccgagag	gcagtagcgt	catattggca	ttaccacggt	ggatgccttg
1501	tcggggaggt	gcttgatgat	gatttccgtg	ttacagggat	caacgcgttg	cgtgtcgttg
1561	atggctccac	attcccttcc	acaccagcga	gccatcctca	gggcttctat	ctgatgttag
1621	ggaggtacat	gggcactaaa	attctgcaag	aaagattagc	ttcagaggag	gctcttcata
1681	agtcaacatt	cgaacccaaa	atcttggagt	cgcttgagtc	agcattatcc	tttgcttttg
1741	aaagttaatt	aataaacccа	tgagttctga	ctcaagttgg	attagggttt	agcttgcaaa
1801	cgggcttcgg	ctctttctat	ccgccatggc	atggtgtacc	tctttatttc	aagtttaata
1861	aattttcctt	tgtgatgttc	tagttctcta	aaataaataa	ataaaataag	gaaaccattg
1921	tgtaaaaaaa	aa				

附录4－3　Swiss－Prot Flatfile 格式

ID　GRAA_HUMAN　　STANDARD;　　PRT;　　262 AA.

AC　P12544;

DT　01－OCT－1989 (Rel. 12, Created)

DT　01－OCT－1989 (Rel. 12, Last sequence update)

DT　28－FEB－2003 (Rel. 41, Last annotation update)

DE　Granzyme A precursor (EC 3.4.21.78) (Cytotoxic T－lymphocyte proteinase

DE 1) (Hanukkah factor) (H factor) (HF) (Granzyme 1) (CTL tryptase)
DE (Fragmentin 1).
GN GZMA OR CTLA3 OR HFSP.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID = 9606;
RN [1]
RP SEQUENCE FROM N. A.
RC TISSUE = T - cell;
RX MEDLINE = 88125000; PubMed = 3257574;
RA Gershenfeld H. K., Hershberger R. J., Shows T. B., Weissman I. L.;
RT "Cloning and chromosomal assignment of a human cDNA encoding a T
RT cell - and natural killer cell - specific trypsin - like serine
RT protease.";
RL Proc. Natl. Acad. Sci. U. S. A. 85:1184 - 1188(1988).
RN [2]
RP SEQUENCE FROM N. A.
RC TISSUE = Blood;
RA Strausberg R.;
RL Submitted (OCT - 2001) to the EMBL/GenBank/DDBJ databases.
RN [3]
RP SEQUENCE OF 1 - 23 FROM N. A.
RA Goralski T. J., Krensky A. M.;
RT "The upstream region of the human granzyme A locus contains both
RT positive and negative transcriptional regulatory elements.";
RL Submitted (NOV - 1995) to the EMBL/GenBank/DDBJ databases.
RN [4]
RP SEQUENCE OF 29 - 53.
RX MEDLINE = 88330824; PubMed = 3047119;
RA Poe M., Bennett C. D., Biddison W. E., Blake J. T., Norton G. P.,
RA Rodkey J. A., Sigal N. H., Turner R. V., Wu J. K., Zweerink H. J.;
RT "Human cytotoxic lymphocyte tryptase. Its purification from granules
RT and the characterization of inhibitor and substrate specificity.";
RL J. Biol. Chem. 263:13215 - 13222(1988).
RN [5]
RP SEQUENCE OF 29 - 40, AND CHARACTERIZATION.
RX MEDLINE = 89009866; PubMed = 3262682;

RA Hameed A., Lowrey D. M., Lichtenheld M., Podack E. R.;
RT "Characterization of three serine esterases isolated from human IL-2
RT activated killer cells.";
RL J. Immunol. 141:3142-3147(1988).
RN [6]
RP SEQUENCE OF 29-39, AND CHARACTERIZATION.
RX MEDLINE=89035468; PubMed=3263427;
RA Krachenbuhl O., Rey C., Jenne D. E., Lanzavecchia A., Groscurth P.,
RA Carrel S., Tschopp J.;
RT "Characterization of granzymes A and B isolated from granules of
RT cloned human cytotoxic T lymphocytes.";
RL J. Immunol. 141:3471-3477(1988).
RN [7]
RP 3D-STRUCTURE MODELING.
RX MEDLINE=89184501; PubMed=3237717;
RA Murphy M. E. P., Moult J., Bleackley R. C., Gershenfeld H.,
RA Weissman I. L., James M. N. G.;
RT "Comparative molecular model building of two serine proteinases from
RT cytotoxic T lymphocytes.";
RL Proteins 4:190-204(1988).
CC -!- FUNCTION: This enzyme is necessary for target cell lysis in cell-
CC mediated immune responses. It cleaves after Lys or Arg. May be
CC involved in apoptosis.
CC -!- CATALYTIC ACTIVITY: Hydrolysis of proteins, including fibronectin,
CC type IV collagen and nucleolin. Preferential cleavage: Arg-|-Xaa,
CC Lys-|-Xaa >> Phe-|-Xaa in small molecule substrates.
CC -!- SUBUNIT: Homodimer; disulfide-linked.
CC -!- SUBCELLULAR LOCATION: Cytoplasmic granules.
CC -!- SIMILARITY: Belongs to peptidase family S1. Granzyme subfamily.
CC ----------

CC This SWISS-PROT entry is copyright. It is produced through a collaboration
CC between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC the European Bioinformatics Institute. There are no restrictions on its
CC use by non-profit institutions as long as its content is in no way
CC modified and this statement is not removed. Usage by and for commercial
CC entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC or send an email to license@isb-sib.ch).

CC -----------------------------------

DR EMBL; M18737; AAA52647.1; -.
DR EMBL; BC015739; AAH15739.1; -.
DR EMBL; U40006; AAD00009.1; -.
DR PIR; A28943; A28943.
DR PIR; A30525; A30525.
DR PIR; A30526; A30526.
DR PIR; A31372; A31372.
DR PDB; 1HF1; 15-OCT-94.
DR MEROPS; S01.135; -.
DR Genew; HGNC:4708; GZMA.
DR MIM; 140050; -.
DR InterPro; IPR001254; Ser_protease_Try.
DR Pfam; PF00089; trypsin; 1.
DR SMART; SM00020; Tryp_SPc; 1.
DR PROSITE; PS50240; TRYPSIN_DOM; 1.
DR PROSITE; PS00134; TRYPSIN_HIS; 1.
DR PROSITE; PS00135; TRYPSIN_SER; 1.
KW Hydrolase; Serine protease; Zymogen; Signal; T-cell; Cytolysis;
KW Apoptosis; 3D-structure.

FT	SIGNAL	1	26	
FT	PROPEP	27	28	ACTIVATION PEPTIDE.
FT	CHAIN	29	262	GRANZYME A.
FT	ACT_SITE	69	69	CHARGE RELAY SYSTEM (BY SIMILARITY).
FT	ACT_SITE	114	114	CHARGE RELAY SYSTEM (BY SIMILARITY).
FT	ACT_SITE	212	212	CHARGE RELAY SYSTEM (BY SIMILARITY).
FT	DISULFID	54	70	BY SIMILARITY.
FT	DISULFID	148	218	BY SIMILARITY.
FT	DISULFID	179	197	BY SIMILARITY.
FT	DISULFID	208	234	BY SIMILARITY.
FT	CARBOHYD	170	170	N-LINKED (GLCNAC...) (POTENTIAL).
FT	STRAND	30	30	
FT	STRAND	33	34	
FT	TURN	37	38	
FT	TURN	41	42	
FT	STRAND	43	48	
FT	TURN	49	51	

FT	STRAND	52	60
FT	TURN	61	62
FT	STRAND	63	66
FT	TURN	68	69
FT	STRAND	76	80
FT	STRAND	84	84
FT	TURN	85	86
FT	TURN	90	91
FT	STRAND	93	102
FT	TURN	104	105
FT	HELIX	108	110
FT	TURN	112	113
FT	STRAND	116	120
FT	STRAND	127	127
FT	TURN	128	129
FT	STRAND	130	130
FT	STRAND	134	134
FT	TURN	138	139
FT	TURN	144	145
FT	STRAND	147	152
FT	STRAND	155	157
FT	TURN	158	159
FT	STRAND	160	162
FT	STRAND	165	165
FT	STRAND	167	173
FT	HELIX	176	180
FT	TURN	181	182
FT	TURN	184	185
FT	TURN	193	194
FT	STRAND	195	199
FT	TURN	201	202
FT	STRAND	206	206
FT	TURN	209	210
FT	TURN	212	213
FT	STRAND	215	218
FT	TURN	219	220
FT	STRAND	221	228
FT	TURN	231	232

FT TURN 234 235
FT TURN 237 238
FT STRAND 241 245
FT TURN 246 249
FT HELIX 252 260
SQ SEQUENCE 262 AA;28968 MW;DA87363A0D92BAF4 CRC64;
MRNSYRFLAS SLSVVVSLLL IPEDVCEKII GGNEVTPHSR PYMVLLSLDR KTICAGALIA
KDWVLTAAHC NLNKRSQVIL GAHSITREEP TKQIMLVKKE FPYPCYDPAT REGDLKLLQL
TEKAKINKYV TILHLPKKGD DVKPGTMCQVAGWGRTHNSA SWSDTLREVN ITIIDRKVCN
DRNHYNFNPV IGMNMVCAGS LRGGRDSCNGDSGSPLLCEG VFRGVTSFGL ENKCGDPRGP
GVYILLSKKH LNWIIMTIKG AV

第5章 DNA和蛋白质的序列生物信息学

序列生物信息学(sequence bioinformatics)是生物信息学的基本内容之一,主要包括两方面的内容:一是序列数据库的建立(参见4.2节);二是DNA和蛋白质的序列分析,主要是序列比对(sequence alignment)和基因结构的分析。基因结构的分析是序列生物信息学和基因组学的重要内容,目前已有两大类基因结构分析软件。一类是根据基因序列预测基因的外显子和内含子的组成,如FGENES,GeneScan等,这类软件已整合到常用基因组数据库中,对新基因进行预测,一般根据HMM(hidden Makov model)算法而建立。另一类是UCSC基因组数据库检索中的BLAT工具,通过已知cDNA或蛋白质序列得到其基因结构和基因组序列。基因结构分析的相关内容在第3章中已作介绍。序列比对分析涉及不同的算法。NCBI数据库中的BLAST,FASTA是最常用的序列比对分析工具,已在4.5节中介绍。本章主要介绍序列比对的常用方法的原理及其应用。

在蛋白质和基因序列分析的早期,人们发现相关蛋白或基因的序列相似,也就是说当把这些序列进行比对时,许多对应的残基相同。这一发现很重要,因为如果两个基因间高度相似,则有足够的理由说明它们之间的同源性。当然,这一点并不总是百分之百的正确。同源(homology)指的是序列来自相同的祖先,或意味着这些序列具有相同的进化历史;而序列的相似(similarity)指的是两序列根据某种参数相像,可以用相同残基的百分比或是其他的方法来表达。相似性不说明任何的历史过程,而是对两个序列间用一定方法进行比较。尽管同源和相似的定义不同,但这两个术语互相关联,因为显著的相似性通常意味着同源。

两个基因间只有同源或不同源的关系。同源的序列被假设起源于一个我们并不知道的共同的祖先序列,通过反复的分子变化而互相有差异。因而,序列比对的目的在于通过同源性和相似位点来推测有关比对序列的进化历史。使用相似性来比较蛋白和基因序列的基础是基于以下的假设:来自同一祖先的序列在进化的过程中逐渐积累随机的

突变，所以分开较久的序列不如分开较近的序列来得相似。因而，序列比对是分析蛋白或基因间进化关系的重要手段。有关进化分析方法参见第 8 章。

序列比较的另一动机是序列的相似往往意味着结构和功能的相似。序列比对中经常可见到某些区域比另外的区域更为保守，而提示这些保守区域对于蛋白的结构和功能更为关键。序列比对的这些特点为所比对的序列提供重要的信息。比如，通过对两个相似功能的蛋白质序列比对的分析可以显示哪一部分序列对于功能最为重要。在那些对于功能不重要的序列区域更容易积累随机的突变，而在功能关键的区域，任何突变都可能使蛋白质失去功能。同样，我们已经知道，如果两个蛋白质序列享有多于 30% 的相同残基，那么它们对应的三维结构基本上总是非常相似，因而，序列比对通常被用做三维结构比对的大致指标。

序列比对可以包含两个或两个以上的序列。对于两个序列的比对称为成对比对（pairwise alignment），而对于多个序列的比较分析则称为多序列比对（multiple sequence alignment，MSA）。

5.1 成对序列比对

如图 5－1 所示，成对序列比对比较两个序列，允许序列间有一定的错配，是计算机在解决生物学问题应用中的一个经典问题。这些问题，如前所述，包括比较两个序列的相似性，从数据库中搜寻相关的序列，提取和比较数据库中的 DNA 和蛋白序列、DNA 序列组装，寻找 DNA 和蛋白质序列中的元件。对于相似性最简单的测量是序列的相同程度，诸如在 DNA 序列组装、相似性搜寻和序列比对过程中部分区域的序列完全相同。但更经常的是，两个比对的序列并不等同，而需要建立非精确的匹配。成对序列比对中所涉及的关键问题包括：①用来评估比对的记分系统；②用来找到最佳比对的运算法则；③不同的比对方式，包括局部比对（local alignment）和整体比对（global alignment）。以下分别阐述几种常用的比对方法。

5.1.1 编辑距离

成对序列比对过程涉及量和质两个方面。量的方面是指用一个数字来测量相似的程度，而质的方面描述的是两个比对的序列在哪里相似和在哪里不同。要理解序列比对的这两个方面，我们需要有一种正式的注解来描述不同程度的相似性。让我们来看以下两个序列："AATTCGAT"和"AATTGGAT"。它们看起来相似，只要把第一个序列中的第 5 位的"C"变成"G"，便得到第二个序列。为确定相似性，很有必要来介绍两个字符串间"距离"的概念。当两个字符串完全相同时，它们间的距离为零，而差异越大，距离越大。确定两个字符串间距离的方法之一是找到从一个序列变成另一个序列所需的变化的量。在上述的例子中，我们需要改变一个字符从第一个字符串得到第二个字符串，因而，我们定义为一个替换（substitution），两个序列间的距离为 1。其他类型的变化包括插入（insertion）和删除（deletion）。当我们向一个序列加入字符时便是插入，而删去字符便是删除。我们也可以把替换看做删除一个字符，接着再插入一个字符。但从计算的角度，把替换定义为一个变化的单位

并无益处。

```
Query:  1    MEKSTMSAILLVLHLFVLLLQYSEVHSLATTSNHDFSYLRFAYDATDLELEGSYDYVIVG   60
             M K   S++L    + +LLL       +     N    +RF  +ATD   E  YDY+IVG
Sbjct:  1    MTKRIDSSLLYTALVVLLLLGVVHRSNARPRVNRPPGFMRFISNATDFASEDYYDYTIVG   60

Query:  61   GGTSGCPLAATLSEKYKVLVLERGSLPTAYPNVLTADGFVYNLQQEDDGKTPVERFVSED  120
             GGT+GCPLAATLS+  ++VL+LERG  +P    PNV++ DGF+  L   ++   +P + F+SE+
Sbjct:  61   GGTAGCPLAATLSQSFRVLLLERGGVPYNRPPNVMSHDGFLTTLTDVNNFDSPAQSFISEE  120

Query:  121  GIDNVRGRVLGGTSMINAGVYARANTSIYSASGVDWDMDLVNKTYEWVEDTIVFKPNYQP  180
             G+ N RGRVLGG+S INAG Y+RA+   +  SG+ WD+  VN++YEWVE  IVF+P  +
Sbjct:  121  GVPNARGRVLGGSSAINAGFYSRADKQFFENSGLVWDLSSVNQSYEWVERAIVFRPQLRT  180

Query:  181  WQSVTGTAFLEAGVDPNHGFSLDHEAGTRITGSTFDNKGTRHAADELLNKGNSNNLRVGV  240
             WQ+    A LE GV P +GF+L+H+ GT+I GSTFD  G RH++ +LL    S+N+RV V
Sbjct:  181  WQTAIRDALLEVGVHPFNGFTLEHKVGTKIGGSTFDRTGRRHSSADLLRYARSSNIRVAV  240

Query:  241  HASVEKIIFSNAP-----GLTATGVIYRDSNGTPHRAFVRSKGEVIVSAGTIGTPQLLLL  295
             +A+VE+++  +++P       ++A GV+YRD  G  H A +R +GEVI+SAG +G+PQLL L
Sbjct:  241  YATVERVLLASSPSVSGSNVSAIGVVYRDQLGRFHHALIRDRGEVILSAGALGSPQLLFL  300
```

图 5-1 成对序列比对

两序列间完全相同的氨基酸残基列在中间行，加号(+)表示不相同的氨基酸阳性替换记分。Query 表示查询序列，Sbjct 表示与之比较的目标序列

根据上述相似性的概念，我们把两个字符串间的编辑距离定义为从一个字符串变成另一个字符串所需要的最小编辑操作，其中包括插入、删除和替换。例如，在比较“AATTGGCG”和“AATCGGC”时，我们可以通过把第一个序列中的第 4 个字母变成“C”，然后删除其后的“G”，便得到第二个序列。如果我们把一个删除、插入和替换都定义为 1，则这两个序列间的编辑距离(edit distance)为 2。

5.1.2 点阵描图

点阵描图(dot plot)是一种简单的图形显示序列相似性的方法。顾名思义，点阵描图是对两个序列间相似程度的视觉表现。沿 X 轴上序列 1 中的每一个单元(核苷酸或氨基酸残基)与沿 Y 轴的第二个序列中的每一个单元进行比较，相同或相似的区域在点阵描图中显示为点组成的对角线。而对角线之外零散的点为背景噪音。所以，如果两个序列在整体范围内都相似，则会在点阵描图中得到一条从一个角到另一对角的对角线。如果两个序列只是具有零散的相似性，那么点阵描图中看到的就是断断续续的对角线。图 5-2 显示的是一个点阵描图的例子，比较 mandelnitrile 水解酶的两个同功酶，MDL1 和 MDL2。在这一计算

中，视窗的长度被设置为20。我们能很清楚地看到一个对角线，表明这两个序列在整体范围具有非常高的序列相似。沿对角线可看到3个大的跳跃或间隔，它们代表两个序列中有插入或删除部位的数目和位置。

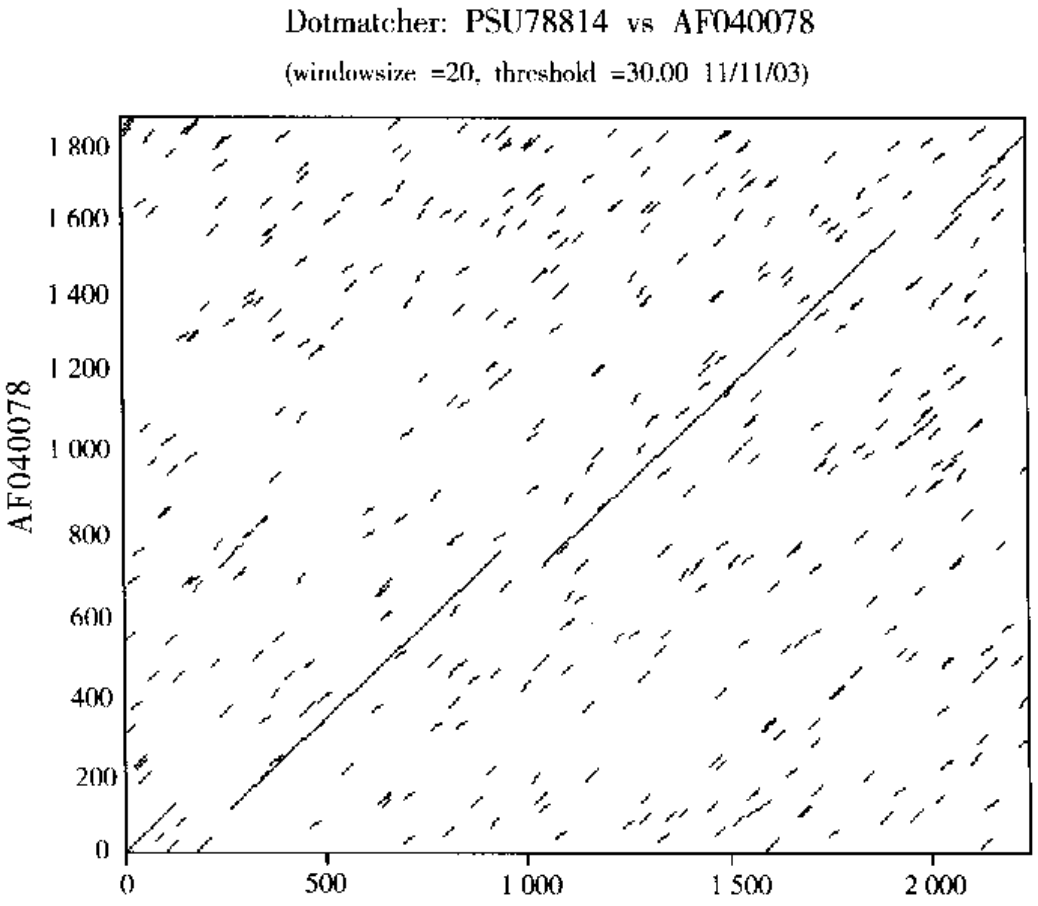


图5-2 序列比较的点阵描图

用Dotmatcher软件(Rice et al.,2000)绘制的MDL同功酶Ⅰ(GenBank PSU78814)和Ⅱ(GenBank AF040078)的氨基酸序列比较点图

尽管点阵描图可能是用于比较两个序列的最古老的方法，但它是一种非常有用的方法。首先，点阵描图不依赖空位(gap)参数，而能很灵敏地找到序列间很微弱的相似。其次，点阵描图不依赖任何先决条件，因而是一种可用于初步分析的理想工具。根据点阵描图的分析结果，用户可以了解所涉及的比对是整体还是局部相似。局部相似意味着在两个从整体看起来并不相似的序列中存在小的相似区域。此外，点阵描图通常允许随时动态地改变最高和最低界限值，可以用来摸索区分信号和背景标准的严格程度。最后，用点阵描图进行同一序列的自身比对，可以发现内部的重复序列或回文序列的基因，蛋白质中重复的结构域(domain)，或是相同残基重复出现的低复杂区(low complexity)。

5.1.3 记分模型

当比较两个序列时，我们在寻找这两个序列通过突变和选择后从同一祖先分化而来的

证据。突变的过程包括替换、插入和删除。后二者比对通常被称为空位。在我们讨论记分模型(scoring model)之前,我们有必要了解一些术语,如:距离(distance)和相似性(similarity)函数、代价(cost)、权重(weight)和记分(score)等。这些不同的名词基本上都属同一类概念,与序列比对数值相关。相似增加比对的分数,而空位降低分数。也就是说,相似越多,分数越高。最佳的比对产生最高的记分。距离函数指的是给错配或空位所记的负分值。尽管相似函数和距离函数的差别在于对数值的解释,但其中的基本逻辑结构相同。在根据整体序列记分的整体比对中,这两个概念是对等的。代价通常是正值,总体的代价应当力求最小。权重和记分既可以是正值,也可以是负值。但经常使用的是用高分来表示高相似。成对比对的过程就是把两个序列上下排列,通过插入不同数目的空位来获得最大数目的相同竖行或列。有两种记分系统用来描述两个被比对序列间的关系。一个是距离的测量,通常用来建立两个相关的生物序列间的关系。通常我们用来估算两个序列间的进化距离,也就是其分别与共同祖先之间的距离。另一记分系统是相似性的测量。距离测量给每一突变记分,两个序列间的距离就是从一个序列转换到另一序列所需所有突变的最小总值,而相似测量根据两序列的相像记分,相似就是两序列相似值的总和。

以下我们将用相似测量来演示记分模型在成对序列比对中的使用。相似测量的基本原理是给相同的字符记正分,而给错配和空位记负分。如果我们用 S 来代表一个相似分数,那么

$$S = \sum (S_{ij}N_{ij}) + \sum (G_kL_k) \qquad \text{(公式 5-1)}$$

其中 S_{ij} 是字符"i" 和字符"j" 配对的得分。例如,一个"G" 核苷酸与一个"G" 核苷酸配对得正分,而一个"G" 与一个"C" 配对得负分。N_{ij} 是字符"i" 和"j" 配对的数目,G_k 是一个长度为 k 的空位的得分,而 L_k 是空位的数目。这种模型允许灵活地记分,但很难知道什么是可以使用的正确数字。通常这一模型被简化,给所有的相同配对记同样的分,给所有的错配和空位记同样的分。其中的 S_{ij} 用常数"s" 或"m" 替换,表示给所有的配对记同样的正分,给所有的错配和空位记同样的负分,而空位的代价分等于空位长度的负值,也就是 $G_k = -k$。

显而易见,上述模型用于 DNA 和蛋白质的序列比对有明显的缺点。首先所有的错配得分相同,这是不理想的。这一问题在蛋白序列的比对中尤其突出,因为某些氨基酸可被不同氨基酸替换的频率上的差别应当在模型中反映出来。其次,空位的代价一般与空位的长度不是线性关系。例如,一个 120 bp 的插入事件不大可能是一个 240 bp 插入事件的 2 倍。我们将先在这里讨论第二个问题,而把第一个问题留到下一节。

为解决上面的第二个问题,我们可以修改以上模型,而使用不同的空位记分方法。已有几种不同的方式来计算修改的空位记分。其中包括 Affine 空位代价、对数空位代价和平方根空位代价。其中 Affine 空位代价使用最广,它包含两个组成部分:

$$G_k = p + (k-1) * e \qquad \text{(公式 5-2)}$$

其中,p 为开放空位的代价,e 为空位延长的惩罚,而 k 为空位的长度。所以一个长度为 1 的空位记分为 p,一个长度为 2 的空位记分为 $p+e$,而长度为 3 的空位记分为 $p+2*e$,如此等

等。一般来说，p 为一个负数，而 e 为一个比 p 小的负数。第二种方式，对数空位代价使用以下的公式：

$$G_k = -[\log(k) + 1] \text{ 或 } G_k = -[\log(k+1)] \qquad \text{（公式 5-3）}$$

对数空位代价试图在空位变长时，降低空位的代价，其根据是当空位延长时，其相对概率的增加应当比长度增加来得慢。第三种方式是平方根（sqrt）空位代价，使用以下或相似的公式：

$$G_k = -\text{sqrt}(k) \text{ 或 } G_k = -[\text{sqrt}(k+1)] \text{ 或 } G_k = -[\text{sqrt}(k)+1] \text{（公式 5-4）}$$

这 3 种方式中，以 Affine 空位代价的记分增加最快，然后是平方根（sqrt）空位代价，而对数空位代价增加最慢。值得指出的是，对于这 3 种空位代价的选择仍有争议。

5.1.4 替换记分

以上所描述的简单记分系统是对相同残基进行记分的一种等同矩阵（identity matrix）。这是用于核苷酸序列分析的方式。但氨基酸记分系统需要增加替换的模式来提高灵敏度以检测弱的相似性。因为已知的事实告诉我们，某些氨基酸很容易被其他生化、物理特性相似的氨基酸所替换。自蛋白序列被发现以来，已有许多不同的记分模式被设立。其中以 PAM 和 BLOSUM 替换记分（substitute score）模式使用最为广泛。理解这些模式如何得来及怎样使用，有助于我们更好地使用序列比对的软件和数据库挖掘（database mining）工具。

由 Dayhoff 和其同事[1]通过研究紧密相似序列间的整体比对，观察氨基酸替换发生的频率而建立的 PAM 记分模式是目前用于氨基酸比对中使用最广的记分模式。其所依据的进化模型是，蛋白质的进化通过一系列在群体中被接受的独立点突变，然后出现在序列库（sequence pool）中。在这一模型中，两个序列间的进化距离被定义为从一个序列演变到另一个序列所需要的最少点突变数目。然而，因为以下的原因，对于两个序列间的净点突变总数并非易事。首先，某一特定的残基有可能先突变，然后通过另外的突变又回复到原来的残基，因而隐藏了突变的效果。其次，特定的位点有可能突变了多次，因而，点突变的实际数目有可能比我们所能看到的要高。

为解决这一问题，Dayhoff 等提出了“被接受的点突变”（accepted point mutation，PAM）（这一术语最初以法语出现，其单词的顺序与英语不同）的概念来定义一个突变为进化过程中基因库中所固定的突变。一个 PAM 单位被定义为每 100 个残基出现一个被接受的点突变。在识别这些被接受的点突变时，Dayhoff 等把他们的分析限于序列的相同百分比在 85% 以上的序列，并且假设氨基酸 X 被氨基酸 Y 替换的可能性等同于氨基酸 Y 被氨基酸 X 所替换的可能性。这样以来，可以构建一个对称的 20 × 20 的矩阵（注：20 指的是氨基酸的种类数目）。这一矩阵的构建从无空位的多序列比对模块（block）开始。模块中每一序列对都足够相似，这一点非常重要。因为 PAM 矩阵的初期目标是建立一个较短进化时区中的过渡矩阵。这一进化时区很短，以致发生多次突变的几率非常低。一个对于包括没有观测到的隐

藏氨基酸变化的突变过程的完整图像可通过计算一种特定氨基酸变化的数目与数据库中同一氨基酸总数之间的比例而得到。这一数据与点突变的数据相结合便得到一个突变几率矩阵(图 5-3)。这样的矩阵适合于特定的进化距离,但也可通过反复自身相乘用来产生适合较大进化距离的矩阵。在 PAM250 矩阵中,一个 PAM 单位指的是序列中 1% 的氨基酸发生突变所代表的进化距离。但这并不意味着 100PAM 单位代表氨基酸组成发生 100% 的变化。因为突变可以反复,因而某些位点上可能发生过多次替换。

A	2																		
R	-2	6																	
N	0	0	2																
D	0	-1	2	4															
C	-2	-4	-4	-5	4														
Q	0	1	1	2	-5	4													
E	0	-1	1	3	-5	2	4												
G	1	-3	0	1	-3	-1	0	5											
H	-1	2	2	1	-3	3	1	-2	6										
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5									
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6								
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5							
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6						
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9					
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6				
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3			
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3		
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

图 5-3 PAM250 记分矩阵

尽管 PAM 及从其演变而来的相关矩阵非常有用,但其中的一些概念,其基础已经受到挑战。比如其中的假设之一认为蛋白质进化速度在序列的各部位上是均一的。现在看来,这一假设很有可能是错误的。因为很明显,蛋白质中保守区的突变率较低而非保守区的突变率较高。Henikoff 等[2]采用了对相差较大序列的多序列局部比对分析,所考虑的替换是发生在蛋白质的相对保守区域,而不是整个序列。相关序列的局部比对可丰富数据库,不同的进化距离的比对通过聚类分析的方式被加入:两个在一定数目的残基上等同的序列被聚集在一起。如果其他更多的序列与序列族中任何序列在长度上等同,将被加入到序列簇中。一簇中的所有序列被平均。每一组序列用来计算所有氨基酸对的替换频率和计算 BLOSUM (blocks substitution matrix)矩阵(图 5-4)。通过改变分开序列簇的极限值可以得到不同的矩阵。例如,BLOSUM80 矩阵是用 80% 的等同极限值得到,而 BLOSUM62 的矩阵是用 62% 的等同极限值获得,如此等等。这样,低值的 BLOSUM 矩阵(如 BLOSUM30)适合进化上高

度分化的序列，而高值的矩阵（如 BLOSUM90）则适合相近的序列。比较而言，BLOSUM62 矩阵与 PAM250 最相似。

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-3	-4	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

图 5-4　BLOSUM62 记分矩阵

值得指出的是没有一种矩阵可以完全适合所有的序列比较。理想的做法是在进行单个的成对序列比较或是数据库序列相似性搜寻时，应如上所述，选择恰当的矩阵。当比较一系列不同距离的蛋白质时，推荐使用 PAM250。这一矩阵在 0～750PAM 范围内提供比其他矩阵更为一致的有效记分。在数据库搜寻中，因为我们并不预先知道序列间的进化关系，当做局部比对时建议使用 PAM40、PAM120 和 PAM250 三种矩阵。低的 PAM 矩阵可以找到短的高相似序列，而高的 PAM 矩阵会找到较长的弱的局部比对。当进行非空位的局部比对时，如非空位的 BLAST 搜寻，BLOSUM62 是最有效的矩阵。

5.1.5　最佳比对

尽管点阵描图，如前所述，可以用来检测序列的相似性，但也有其缺点，如随机配对造成的背景噪音、对角线不易被计算机检测，也易被目测所疏忽。点阵描图不能解决被低相似区或插入/删除所间隔的序列之间的相似性。因此，需要一种通过点阵矩阵找到一条路径来提供两序列间的最佳比对方法。这一方法就是最佳比对（optimal alignment）。

最佳比对的主要目的是使用某种模型从所有可能的比对中挑选具有最高记分的比对。但在序列比对中,并没有什么真正的最佳比对,因为最佳性总是依赖于建立在比对上的假设。例如,到底空位应当得到多少罚分?所有的序列比对都需对此作相应的假设。让我们来看以下一个极简单的例子来比对两个 DNA 序列:TCCG 和 TTACG。以下显示几种其可能的比对:

```
1          2          3          4          5
TC.CG      T.CCG      .TCCG      T..CCG     .T.CCG
|  | |     |   | |      | | |     |  |  |      |  |  |
TTACG      TTACG      TTACG      TTAC.G     TTAC.G
```

根据所选的模型不同,这两个序列间的最佳比对不止一种。如果使用的记分模型为:等同配对 $s=1$,错配 $m=-1$ 和空位 $g=-1$,以上比对的记分为:

比对方式	1	2	3	4	5
记分	1	1	1	0	0

如果使用的记分模型为:等同配对 $s=2$,错配 $m=-2$ 和空位 $g=-1$,以上比对的记分为:

比对方式	1	2	3	4	5
记分	3	3	3	3	3

如果使用的记分模型为:等同配对 $s=3$,错配 $m=-3$ 和空位 $g=-1$,以上比对的记分为:

比对方式	1	2	3	4	5
记分	5	5	5	6	6

当序列长度增加时,可能带有空位的比对数目通常非常大。我们如何能找到“最好”的呢?手工比对不是没有可能,但非常烦琐,而且尝试每一种可能空位的方法非常耗时,大致等于 N^M。其中 N 和 M 为两个序列的长度。显然,我们不可能对所产生的所有可能的比对来计算其记分。即便是较短的序列,可能的比对数目也是超出可及的范围。因此,我们只能使用一种运算法则来找到一个或一组最佳比对。这种类型的运算法则就称动态编程(dynamic programming)。

动态编程是一种用于解决某些优化问题的常用运算法则。“编程”一词在此并不是计算程序,而是一种在数学上用一套规则来得到答案的行话。动态编程尝试通过把问题分成许多小问题来解决。它使用的是一种“由下而上”(bottom-up)的方式,也就是说小问题被排列,用一种系统的方式得到解决,而最终解决原始的问题。这种由下而上的方式比与其相对的“由上而下”(top-down)的方式更为有效,因为避免了同样问题的重复。动态编程一般包括两个步骤:①建立一种递归式的策略;②建立一种不用递归的“由下而上”的策略。动

态编程是计算序列分析中的中心环节,可用于不同方面。以下我们讨论其在整体和局部序列比对中的应用。

5.1.5.1 Needleman – Wusch 整体比对运算法则

整体比对从整个序列的范围比较两个序列,适合于比较在整体范围内具有相似性序列的比较。最早和著名的运算法则为 Needleman – Wusch 运算法则[3],使用动态编程来找到最佳序列比对。让我们以两个序列 X 和 Y 的比对为例子。既然允许有空位,这两个序列的长度不一定要求相同:

$$X: x_1x_2\cdots x_n$$
$$Y: y_1y_2\cdots y_m$$

$x_1x_2\cdots x_n$和 $y_1y_2\cdots y_m$间的最佳比对可以有以下 3 种可能:

1 x_i与 y_i配对
2 x_i与一个空位配对
3 y_i与一个空位配对

假设 $F(i, j)$等于亚序列 $x_1x_2\cdots x_i$和 $y_1y_2\cdots y_j$最佳比对的记分。我们把这些记分排在一个 $n \times m$ 的矩阵 F,然后将其用动态编程来填满。如果使用线性的空位罚分 $=d$,F 矩阵从 $F(0,0) = 0$ 开始填起,第一横行从 $F(i,0) = -id$ 开始,而第一竖行从 $F(0,j) = -jd$ 开始。也就是:

1 $F(0, 0) = 0$
2 i 为 1 到 m 时, $F(i, 0) = -i \times d$
3 j 为 1 到 n 时,$F(0, j) = -j \times d$

矩阵的其余部分从左上到右下角用以下的公式填满:

$$F(i,j) = \max\begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases} \quad \text{(公式 5-5)}$$

每一个 $F(i, j)$ 位置上的记分建立在 3 种可能的前一位置:

$$\begin{array}{ccc} F(i-1, j-1) & & F(i-1, j) \\ & {\scriptstyle +s(x_i, y_j)} \searrow \quad \downarrow {\scriptstyle -d} & \\ F(i-j-1) & \xrightarrow[-d]{} & \boxed{F(i, j)} \end{array} \quad \text{(公式 5-6)}$$

这样我们可以把分数的计算一行行从上到下，一列列从左到右，一直到矩阵的最后一个位置 $F(n,m)$，从而得到序列 $x_1x_2\cdots x_n$ 和 $y_1y_2\cdots y_m$ 之间最佳比对的记分。

要得到真正的最佳比对，而不仅是其记分，动态编程表中的指针记录如何得到每一步。也就是说，对每一个位置 $F(i,j)$，都有一个指向前一位置 $F(i-1,j-1)$、$F(i-1,j)$ 或 $F(i,j-1)$ 的指针。从 $F(n,m)$ 到 $F(0,0)$ 的整个路径可以通过这些指针来记录。其中，一个在对角 $F(i,j)$ 上的指针标志着 x_i 和 y_j 的配对，朝上的指针标志 x_i 和一个空位的配对，向左的指针指向 y_i 和一个空位的配对。图 5-5 显示使用配对 $m=1$、错配 $s=-1$ 和空位罚分 $d=-1$ 的记分模型来为序列 X = ACAATCGTTA 和 Y = CAAACAAGTT 比对的矩阵 F 的建立方法，并演示指针在计算的过程中是如何用来记录最佳的比对。

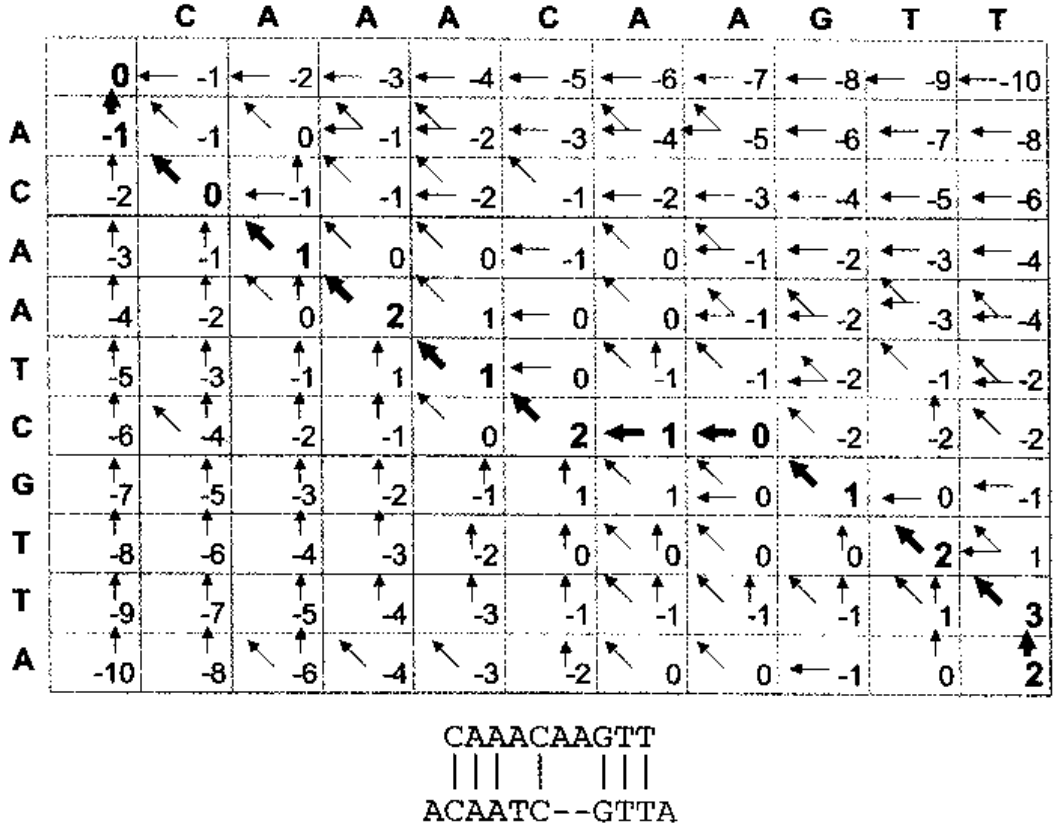


图 5-5 整体动态编程矩阵 F 示意图

配对 $m=1$，错配 $s=-1$，空位 $d=-1$（上方）；最佳整体比对记分为 2（下方）

5.1.5.2 局部比对的 Smith-Waterman 运算法则

我们上面提到，整体比对的算法对序列的整体范围进行比对。然而，在许多情况下，序列间只有有限的相似性，可以是一个共有的结构域或一段可识别的相似区域。例如，多结构域蛋白间可能只有其中一个结构域相同，而其余的结构域不同，或是有重复区域的序列。这些情形一般用 Smith-Waterman 运算法则来处理[4]。局部比对的目的是从每一个序列中找到一个区域来配对，用做这两个序列间的最佳比对。这一算法依赖于一个使记分最大化的记分方法，要不然空的比对总是产生最小的距离。这种记分方式与整体比对的记分方式非常相似。与 Needleman-Wunsch 算法相比较，Smith-Waterman 运算法则的关键点在于如下所示的矩阵中的每一位点通过第 4 个可能性“0”来标示一个新的比对的开始。

$$F(i,j) = \max\begin{cases}0\\ F(i-1,j-1) + s(x_i,y_j)\\ F(i-1,j) - d\\ F(i,j-1) - d\end{cases} \qquad \text{(公式 5-7)}$$

其他的修改包括把第 -行和第 -列的起始值定为零。也就是：

1 $F(0,0) = 0$
2 i 为 1 到 m 时，$F(i,0) = 0$
3 j 为 1 到 n 时，$F(0,j) = 0$

在用此方式通过动态编程将矩阵填满后，对路径的回瞻从矩阵中数值最大的位点开始，这一点不一定是 $F(n,m)$，并且回瞻到遇到一个值为零的位点时停止，或一直到 $F(0,0)$。图 5-6 中显示的就是这一过程。其中与序列 X = ACAATCGTTA 和 Y = CAAACAAGTT 比对相关的 F 矩阵的构建使用的是配对 $m = 1$、错配 $s = -1$ 和空位罚分 $d = -1$ 的记分模型。图中显示的是这两个序列的最佳局部比对。

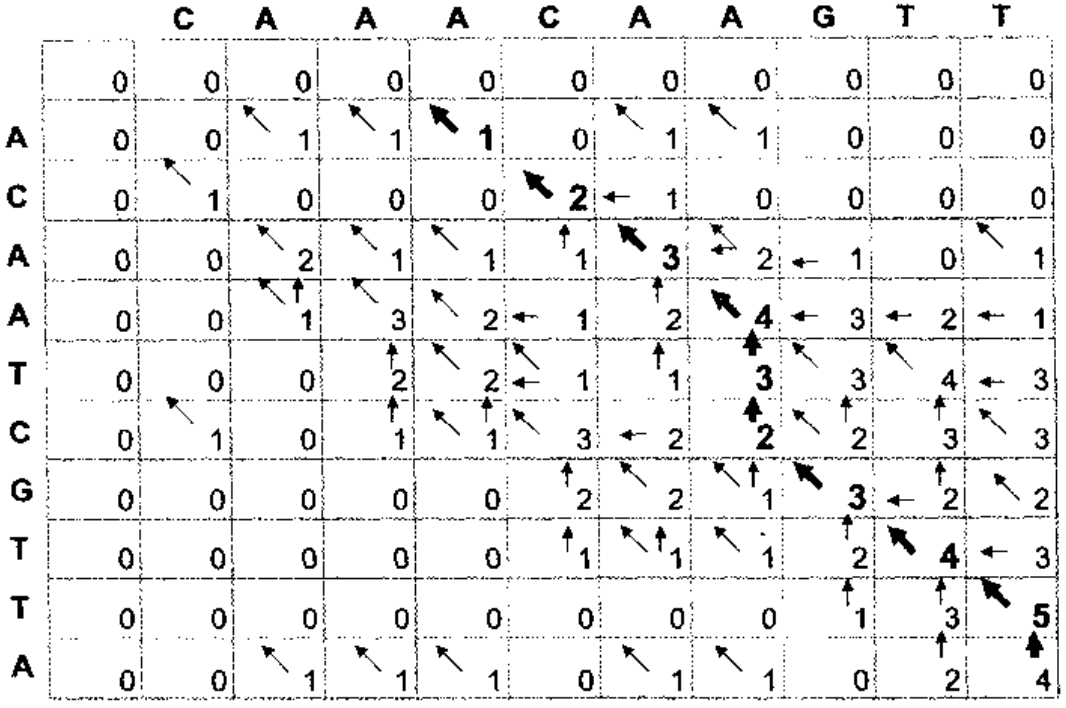


图 5-6　局部动态编程矩阵 F 示意图

配对 $m=1$，错配 $s=-1$，空位 $d=-1$（上方）；最佳整体比对记分为 5（下方）

（胡自华著　梁　平译）

5.2 多序列比对

5.2.1 多序列比对主要方法和常用软件工具

除开常用的成对序列比对之外，生物信息学研究中也经常需要对多于两个以上的序列进行同时比对。比如，从一个 BLAST 搜寻往往得到大量的相似序列，要理解这些序列之间的相互关系或是找出为所有序列所共有的区域，方法之一是把所有的序列逐一进行两两成对比对。但更为有效的途径是能对这些序列同时进行比对，并且把比对的结果放在一起。这一过程称为多序列比对（multiple sequence alignment，MSA）。多序列比对可用于 DNA 序列的分析，以寻找个体之间单核苷酸多态性（SNPs）、外显子和基因调节因子或是研究相近物种之间的进化关系。但多序列比对更多的是用于蛋白质序列，来研究序列之间在初级序列和高级结构上的相似性，搜寻保守的结构域或基序，以及研究物种间的进化关系（参见第 8 章）。表 5－1 所列为目前 MSA 中常用的方法和相应软件工具。

表 5－1 用于 MSA 的常用软件工具的网站

名称	来源
整体比对	
CLUSTALW/X	下载：ftp://ftp.ebi.ac.uk/pub/software
	Online server：http://www.ebi.ac.uk/clustalw/
MSA	网址：http://www.psc.edu
	网址：http://www.ibc.wustl.edu/ibc/msa.html
PILEUP	下载：ftp://fastlink.nih.gov/pub/msa
PRALINE	网址：http://ibivu.cs.vu.nl/programs/pralinewww/
DIALIGN	网址：http://www.genomatix.de/cgi－bin/dialign/dialign.pl
MultAlin	网址：http://prodes.toulouse.inra.fr/multalin/multalin.html
PRRP	下载：ftp://ftp.genome.ad.jp/pub/db/hgc/software/saitama－cc/
SAGA	下载：http://igs－server.cnrs－mrs.fr/~cnotred/_Projects_home_page/saga_home_page.html
蛋白质局部比对	
Asset	下载：ftp.ncbi.nlm.nih.gov/pub/neuwald/asset
BLOCKS	网址：http://blocks.fhcrc.org/blocks
eMOTIF	网址：http://dna.stanford.edu/emotif/
GIBBS	下载：ftp.ncbi.nlm.nih.gov/pub/neuwald/gibbs9_95/
HMMER	下载：http://hmmer.wustl.edu/
MACAW	下载：ftp.ncbi.nlm.nih.gov/pub/macaw/
MEME	网址：http://meme.sdsc.edu/meme/website/
SAM HMM	下载：http://www.cse.ucsc.edu/research/compbio/sam.html

5.2.2 渐进整体比对法

5.2.2.1 CLUSTALW

CLUSTAL 是用于 MSA 的最为流行的软件工具，为此，本节将作较为详细的介绍，而对其他类似方法则只作简要描述。CLUSTAL 最初由 Higgins 等于 1988 年创立并不断完善[5,6]。CLUSTALW 和 CLUSTALX 为其最新的版本。版本 W 和 X 核心功能完全相同，但 X 版具备图形功能，而 W 版保留以前的非图形平台。CLUSTAL 有用于 PC 和 UNIX/LINUX 平台的各种版本。CLUSTAL 实施的是整体序列比对，方法与 MSA 程序不同。其步骤是：①进行所有序列间的成对比对；②基于成对比对分数产生一个相邻连接（参考下面的章节）进化树（neighbor－join tree）；③根据进化树提供的序列间关系按顺序对序列进行比对。先排最相近的序列，再添加其他的序列产生一个 MSA 来显示每列的序列变异。初期的比对可通过一种与 FASTA 相似的较快的 k－tuple 模型寻找方法或是通过完全的动态编程方法。其中的优化比对分数由强化的动态编程完成。在产生进化树的过程中需要计算序列间的遗传距离（参见第 8 章）。遗传距离就是错对位置的数目除以比对位置的总数（空位除外）。

MSA 对空位（gap）的记分方法与成对比对中的方法不同。当新的序列加入到已有的 MSA 中时，新积累的空位会影响到之后序列的比对。根据 Pascarella 和 Argos 的研究表明，结构相近的蛋白质序列相比对时，空位有选择性地位于二级结构单元之间。根据他们所列的与这些区域中每一种氨基酸相邻的空位频率表，CLUSTAL 使用一种独特的方法把空位放置在保守的结构域之间。与其他比对程序一样，CLUSTAL 对空位的出现和延伸计负分。用户可以调整负分的多少。CLUSTAL 的使用相对简单，W 版使用菜单形式。表 5－2 举例说明每级菜单的使用。第一步是序列的输入。第一级菜单为主要功能的选择（表 5－2A）。二级菜单提供一级菜单上每一功能中的参数和其他选项。表 5－2B 显示 MSA 的产生方法。除非序列数目非常多并且相当相似，一般应选择慢速而精确的比对方法。从这一级菜单可以进入到修改 MSA 参数的第三级菜单（表 5－2C）。有经验的用户可根据具体的情况调整包括空位罚分和不同蛋白质权重模式等多项参数。对于包含差异较大的序列在内的比对，用户可以先使用缺损参数，研究其输出结果。如果不满意，可尝试不同的参数，观测比对结果的改变而选择最佳的配套参数。用户如果需要对输出的 MSA 结果加上较为易看、美观的框架阴影（BOXSHADE），可将结果剪接和拷贝到以下网页：http://www.ch.embnet.org/software/BOX_form.html，获得包括 jpg、rtf、ps、EPS 等多种格式。

表5-2A CLUSTALW 一级菜单

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
        CLUSTAL W (1.83) Multiple Sequence Alignments
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

    1. Sequence Input From Disc 从驱动器上输入序列
    2. Multiple Alignments 多序列比对
    3. Profile / Structure Alignments ? /结构比对
    4. Phylogenetic trees 进化树
    S. Execute a system command 执行系统命令
    H. HELP 帮助
    X. EXIT (leave program) 退出

    Your choice: 选项
```

表5-2B CLUSTALW 二级菜单(MSA)

```
* * * * * * MULTIPLE ALIGNMENT MENU * * * * * *

    1. Do complete multiple alignment now (Slow/Accurate) 慢速但精确的完全 MSA
    2. Produce guide tree file only 只产生引导树文件
    3. Do alignment using old guide tree file 使用老的引导树进行比对
    4. Toggle Slow/Fast pairwise alignments = SLOW 慢速/快速成对比对反复开关
    5. Pairwise alignment parameters 成对比对参数
    6. Multiple alignment parameters 多序列比对参数
    7. Reset gaps before alignment? = OFF 比对前重设空位
    8. Toggle screen display      = ON 屏幕显示反复开关(defaul:开)
    9. Output format options 输出格式选项
    S. Execute a system command 执行系统命令
    H. HELP 帮助
    or press [RETURN] to go back to main menu 按回车回到主目录

Your choice: 选项
```

表 5-2C CLUSTALW 三级菜单(MSA 参数)

```
* * * * * * * * * MULTIPLE ALIGNMENT PARAMETERS * * * * * * * * *

1. Gap Opening Penalty 打开空位罚分                    :10.00
2. Gap Extension Penalty 空位延伸罚分                  :0.20
3. Delay divergent sequences 延迟差异序列              :30 %
4. DNA Transitions Weight DNA 过渡权重                 :0.50
5. Protein weight matrix 蛋白质权重模式                :Gonnet series
6. DNA weight matrix DNA 权重模式                      :IUB
7. Use negative matrix 使用负模式                      :OFF
8. Protein Gap Parameters 蛋白质空位参数

H. HELP 帮助

Enter number (or [RETURN] to exit): 输入数字或回车退出
```

5.2.2.2 PILEUP

PILEUP 是 GCG 软件包中的 MSA 分析工具,使用的方法与 CLUSTAL 非常相似。先使用 Needleman - Wunsch 动态编程所有序列进行成对比对,根据其记分按 UPGMA 的方法产生一个进化树。以此进化树来引导最相近序列和成组序列的比对。比对过程中使用标准的记分和罚分矩阵。与 CLUSTAL 相比,其主要缺点是缺乏任何新近的在罚分和记分方面的改进。

CLUSTAL 和 PILEUP 都使用渐进式整体比对方法(progressive global alignment),其缺点是最终的比对结果取决于最初的成对比对。因此,如果起初的序列相似程度高,则比对准确,那么初期的错误也较少。但如果起初比对的序列亲缘关系较远,比对的错误会较多,而这些错误会在后期的 MSA 中进一步放大。另外的问题就是所选的记分和罚分的标准并不一定适合一组序列中的所有序列。因此,其结果不一定总是准确无误。用户有时需要对所得的 MSA 进行适当的手工调整。

5.2.3 反复整体比对法

为克服渐进整体比对的以上问题,便产生了反复比对法(iterative alignment)。其策略是在比对的过程中不断地重新比对各亚组序列,然后把亚组序列再排成包括所有序列在内的整体比对。其目标是要得到一个最大的由成对比对分数相加而成的总比对分数。对于亚组的选择基于通过与相似方法产生的进化树中序列的顺序。MultAlin 程序在渐进比对的过程中重新计算成对比对的分数,调整用来完善比对记分的进化树。DIALIGN 程序先通过在成对比对中采用对角点阵描图来找到不包含空位的局部比对区域,然后找到一系列给予最大总和的加权过的对角比对。

5.2.4 遗传算法

遗传算法(genetic algorithm)是一种由计算机科学家发明的普通机器学习算法。这一方法不久前由 Notredame 和 Higgins 应用于 SAGA(sequence alignment by genetic algorithm)软件包。这一方法背后的基本原理是通过重新安排模拟进化过程中空位的插入和重组来尝试许多的 MSA,以达到越来越高的 MSA 记分。这一方法的缺点是当序列的数目超过 20 时,其速度变得非常慢。另一相似的方法是通过一种称为模拟退火的概率途径来调整已有的比对来获得高记分的 MSA。MSASA 程序(multiple sequence alignment by simulated annealing)从一个经验的 MSA 开始,然后根据一个确定增加比对分数的算法来调整比对。遗传算法的成功可能归功于其对序列的调整,因为有理由相信这类事件在蛋白家族的进化过程中确实会发生。

5.2.5 局部比对

以上所讨论到的 MSA 方法是序列的整体比对,即包含序列的所有部分。局部比对(local alignment)方法能够确定序列中高度保守的区域,并产生一种称为剖面(profile)的记分矩阵来记录保守区内每一行中氨基酸替换的记分。另一种做法是比对过程中只扫描称为模块(block)的没有空位的区域。还有第三种方法即不首先产生比对,而是通过模型搜寻的统计方法来从一组序列中寻找相似序列区域。以下简要地讨论局部比对的主要方法。

5.2.5.1 剖面分析

剖面分析通过对一组序列进行整体 MSA 分析,然后把其中高度保守的区域提出,分成小的 MSA。从这些 MSA 根据其序列和结构的比对得到一个记分矩阵,列出每一位置上对不同残基的记分,称为位置特异的记分表(position - specific scoring table)或叫剖面。这一记分方法不仅融入了 Dayhoff 的突变距离矩阵,也代表了对相似序列组成的蛋白质家族结构研究的结果。图 5 - 7 显示的就是根据一组来自于不同生物的 HSP70 蛋白所建立的剖面。图中左边第一列为共识序列(consensus sequence)"ITLSTTCVCV"。其余行的数值表示共识序列中每一位置被每一种其他的氨基酸或空位替换的几率,具体为在实际序列中每一氨基酸出现的频率的对数与随机频率对数的比例。如果出现空位,则必须减去空位行在相应位置上的数值。

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len
I	8	3	-2	5	4	5	5	-4	24	0	15	13	1	1	1	-7	2	22	21	-18	-6	4	100	100
T	13	19	-5	24	18	-18	19	7	1	7	-7	-4	14	11	10	-1	9	29	3	-28	-14	15	100	100
L	5	5	-5	3	4	13	4	2	8	-4	14	12	8	-5	0	-10	0	10	10	-1	5	2	22	22
S	17	14	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	34	19	1	-8	-15	4	100	100
T	15	3	22	0	-1	-5	12	-2	-7	-3	-8	-6	5	7	-8	-7	16	29	9	-22	6	-4	100	100
T	8	1	12	2	0	5	6	-4	19	-4	8	5	-1	2	-8	-8	7	22	19	-15	4	-3	100	100
C	17	0	24	-1	-3	11	8	-1	7	-10	1	-2	1	-3	-8	-14	8	5	9	-5	14	-7	100	100
V	11	0	18	-1	-2	2	14	-10	26	-4	9	7	-3	7	-7	-7	21	10	31	-19	-5	-5	100	100
C	10	-8	15	-11	-11	6	8	-7	11	-10	4	3	-7	0	-11	-4	11	5	15	-22	14	-11	100	100
V	7	7	-3	8	8	-3	11	1	20	-1	14	10	4	2	8	-5	0	5	26	-24	-6	8	100	100

图 5 - 7 对一组 HSP70 蛋白质的剖面分析

任何其他蛋白序列与此组蛋白的相似性可以通过与其剖面的比较用动态编程的方法来测试。剖面与常用的序列比较方法有两个主要的不同点:①可以用任何数目的已知序列来建立剖面,因而允许使用比成对比对方法更多的信息;②剖面中包括了对每一位置上的插入或删除的罚分,而允许考虑到序列的二级结构。用于剖面分析的软件工具之一——PROFILE - SS 可以从http://www.psc.edu/general/software/packages/profiles/得到。剖面分析的缺点是一个剖面仅代表 MSA 中所包括序列中所存在的变异。如果其中的几个序列特别相似,那么所得的剖面就会偏向这些序列。

5.2.5.2 模块分析

与剖面分析相似,模块也代表 MSA 中的保守区。但不同的是,模块没有插入和删除,每一个位置上只有配对和错配,因而每一个序列中的同一模块具有相同长度。模块的寻找可以通过计算机程序,如 MOTIF 程序,或是统计方法,如 ASSET。也可以从 MSA 中提取。图 5 - 8 显示从一组蛋白序列中所找到的保守区,即模块。模块数据库中包含从多达 400 个序列所组成的 MSA 中提取的宽度为 10 ~ 55 残基的模块。用户可以把自己的序列以 FASTA、CLUSTAL 或 MSA 的格式发送到其网站而获得用户序列中所包含的已知模式或基序。

```
rhle_ecoli    GVDVLVATPG  RLLDLEHQNA  · · · · VKLDQV  EILVLDEADR  MLDMGFIHDI
dbp2_schpo    GVEICIATPG  RLLDMLDSNK  · · · · TNLRRV  TYLVLDEADR  MLDMGFEPQI
dbp2_yeast    GSEIVIATPG  RLIDMLEIGK  · · · · TNLKRV  TYLVLDRADR  MLDMGFEPQI
dbpa_ecoli    APHIIVATPG  RLLDHLQKGT  · · · · VSLDAL  NTLVMDEADR  MLDMGFSDAI
rm62_drome    GCEIVIATPG  RLIDFLSAGS  · · · · TNLKRC  TYLVLDEADR  MLDMGFEPQI
 p68_human    GVEICIATPG  RLIDFLECGK  · · · · TNLRRT  TYLVLDEADR  MLDMGFEPQI
rhlb_ecoli    GVDILIGTTG  RLIDYAKQNH  · · · · INLGAI  QVVVLDEADR  MYDLGFIKDI
yn21_caeel    RPHIIVATPG  RLVDHLENTK  · · · GFNLKAL  KFLIMDEADR  ILNMDFEVEL
yhm5_yeast    KPHIIIATPG  RLMDHLENTK  · · · GFSLRKL  KFLVMDEADR  LLDMEFGPVL
me31_drome    KVQLIIATPG  RILDLMDKKV  · · · · ADMSHC  RILVLDEADK  LLSLDFQGML
drs1_yeast    RPDIVIATPG  RFIDHIRNSA  · · · SFNVDSV  EILVMDEADR  MLEEGFQDEL
if4a_rabit    APHIIVGTPG  RVFDMLNRRY  · · · · LSPKYI  KMFVLDEADE  MLSRGFKDQI
if41_human    APHIIVGTPG  RVFDMLNRRY  · · · · LSPKYI  KMFVLDEADE  MLSRGFKDQI
vasa_drome    GCHVVIATPG  RLLDFVDRTF  · · · · ITFEDT  RFVVLDEADR  MLDMGFSEDM
srmb_ecoli    NQDIVVATTG  RLLQYIKEEN  · · · · FDCRAV  ETLILDEADR  MLDMGFAQDI
DEAD_ecoli    GPQIVVGTPG  RLLDHLKRGT  · · · · LDLSKL  SGLVLDEADE  MLRMGFIEDV
if4a_orysa    GVHVVVGTPG  RVFDMLRRQS  · · · · LRPDYI  KMFVLDEADE  MLSRGFKDQI
DEAD_klepn    GPQIVVGTPG  RLLDHLKRGT  · · · · LDLSKL  SGLVLDEADE  MLRMGFIEDV
p110_mouse    GCHLLVATPG  RLVDMMERGK  · · · · IGLDFC  KYLVLDEADR  MLDMGFEPQI
 p54_human    TVHVVIATPG  RILDLIKKGV  · · · · AKVDHV  QMIVLDEADK  LLSQDFVQIM
```

图 5 - 8 从一组蛋白的 MSA 中提取保守的模块

5.2.6 序列比对中的统计方法

5.2.6.1 最大期望运算法则(expectation maximization algorithm,EM)

EM 已被用来从无法比对的蛋白序列中寻找保守的区域或从 DNA 序列中找到 DNA 结

合位点[7]。这些基序(motif)中允许空位的存在。此法的步骤是首先对基序所在每一个序列中的位置和大小进行一个大致的猜测,并将序列中的这些部分进行比对。这一比对提供基序中每一位置上的残基或核苷酸组成的大致估计。之后,此法使用两个依次反复重复的步骤。第一步称为期望步骤:从上述已有的基序中通过每列中的组成来估算每一序列的每一位置上找到这一位点的概率。这些概率反过来用来更新位点中每一列上残基或核苷酸的组成。第二步为最大化步骤:使用以上所得数据重复上面的步骤,直到达到每列的组成不再改变。使用上述方法的一个网站工具为 MEME,为加州大学 San Diego 分校的超级计算中心所创立。这个工具有两个版本,即 MEME (multiple EM for motif elicitation) 和 ParaMEME (http://www.sdsc.edu/MEME/meme/website.html)。

5.2.6.2 Gibbs 抽样(Gibbs sampler)

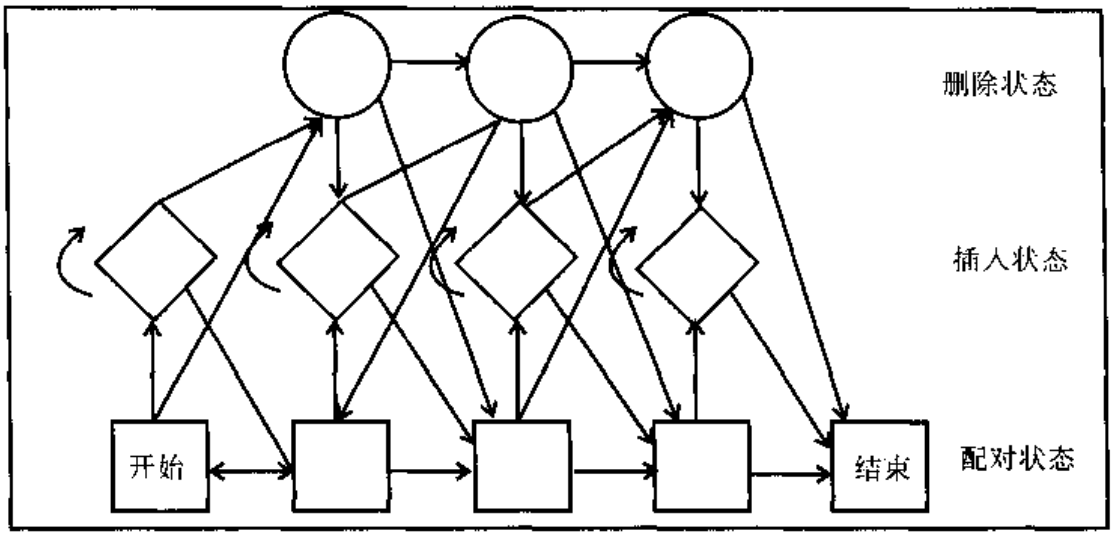
这一方法的基本原理与 EM 相似,是通过统计方法从一组序列中找到可能存在的基序,确定其最佳的长度和在每个序列出现的次数。但其算法与 EM 不同。Gibbs 抽样包括两个步骤。第一步随机地从所有序列中抽出一个或几个作为外围序列用来计算背景噪音。然后对剩下的序列随机地选择假设存在的基序的开始位置,并来回移动其位置来寻找最可能为所有序列所共有的基序。获得这个基序的概率必须高于背景噪音。第二步试图调整序列比对以获得较好的记分,但保留寻找其他更好位置的余地。当在几个序列中选择了基序的正确开始位置后,基序的组成便开始反映在其余序列中所能找到的基序。最后是集合最佳的基序,并确定基序在每一序列中的精确位置。这一方法被成功地用做从几乎没有相似性的序列中找到 30 个螺旋转螺旋(helix-tura-helix)DNA 结合结构域和其他多个难度较大情形下基序的寻找。因而,此法被广泛用于从蛋白序列中发现复杂和可变的基序。

5.2.6.3 HMM 模型(Hidden Markov Model, HMM)

HMM 是一种统计模型,在生物信息学中已被用来产生序列剖面、预测基因结构和蛋白结构。其中有名的 Pfam 蛋白家族数据库就是使用 HMM 方法所建立。除应用于序列分析之外,HMM 已被成功地应用于语音的识别。在对一组序列进行比对时,它试图考虑配对、错配和空位的所有可能结合情形(图 5-9)。HMM 分析从一组已知的相关序列开始。这一组序列包含 20 个或更多的序列,它们用来训练 HMM 模型。HMM 的优点是以很强的概率理论为基础,无须对所分析的序列排序,无须考虑插入和删除的罚分及可以利用实验数据信息。其局限性是对于新剖面的分析必须要有以一组包括至少 20 个已知的序列作为先决条件。训练过程中包括的序列越多,其分析的精确性越高。序列分析中运用 HMM 模型的程序有 Krogh 等[8]建立的 SAM 和 Eddy[9]创立的 HMMER 软件。

5.2.6.4 位置特异性记分矩阵(position-specific scoring matrix, PSSM)

对于序列保守区的 MSA 分析可以产生位置特异的记分矩阵 PSSM。PSSM 的表示方法与图 5-7 中所显示的剖面相似,都是用一个数值表来表示每一个氨基酸或核苷酸在每一位置上出现的频率。只是 PSSM 直接使用频率的对数值($\log_2$)为其数值。PSSM 可用来搜寻用户序列以确定 PSSM 所代表的序列模式的最可能位置。另外,PSSM 可用来搜寻整个数据库以寻找额外的具有相同模式的序列。PSSM 不仅可以用于蛋白质家族所共有的序列模式,也被用于寻找 DNA 序列中与基因调节位点、内含子与外显子交界区等相关的序列模式。



(Krogh, 1998)

图 5-9 HMM 模型示意图

5.2.6.5 序列标语(sequence logo)

序列标语是一种用图形来表示基序中每一列的残基信息的方法。其中的序列信息从前面所描述过的 PSSM 衍生而来。蛋白质和 DNA 序列中的保守模式都可以用序列标语来表示。图 5-10 所表示的是 Alu 插入位点序列标语 AAA/GA(作者未发表数据)。X 轴为模式中的位置,而 Y 轴上字母的高度代表该字母在此位置上出现的频率。用序列标语表示的模块可以从模块网站上获得。另外用户也可到 http://weblogo.berkeley.edu,根据自己的数据绘制序列标语。

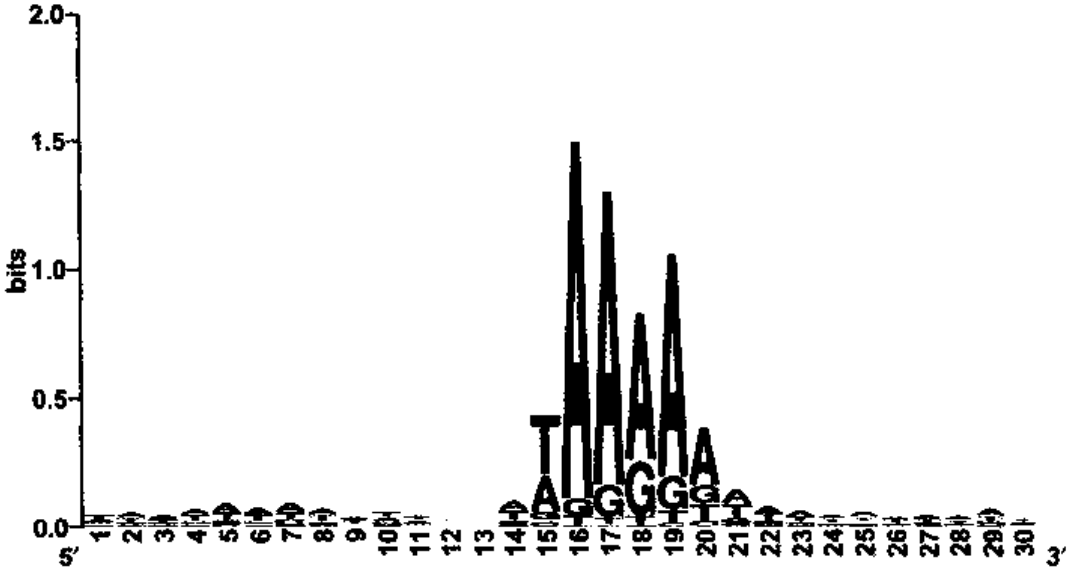


图 5-10 序列标语分析图

5.2.7 MSA 的编辑和修改

一旦用上述方法之一获得 MSA，有可能需要通过手工编辑来得到更为合理或期望中的比对。在选择适合的编辑器时，应当尽可能考虑以下因素：①能利用彩色显示器来显示用不同颜色代表的氨基酸来获得比对的清晰视觉效果；②能够识别用各种程序产生的 MSA 格式，并能在编辑后保持原有的格式；③有合适的视窗界面，能利用鼠标器进行序列的添加、删除和挪动。以下是几种常用的 MSA 编辑软件。

（1）CINEMA（colour interactive editor for multiple sequence alignment） 这是一种广泛使用的序列和点阵图谱编辑工具，由 A. W. R. Payne 等编写，使用在万维网浏览器中运行的 Applet，因而可以在任何平台上使用。CINEMA 网页为http://bioinf. man. ac. uk/dbbrowser/CINEMA2. 1/。下载地址为ftp://ftp. bioinf. man. ac. uk/pub/CINEMA/2. 1/CINEMA2. 1 - default Config. tar. gz。其特点包括拖放功能、序列左移或右移、使用分裂视窗来显示比对的不同区域、多基序的选择和操作等多种功能。

（2）GDE（genetic data environment） 这是一种用于 Unix 机器的序列比对编辑和显示工具。可从ftp. ebi. ac. uk/pub/software/unix/无记名下载。GDE 的功能已被加入到 GCC 软件包中。如果用户在终端机上有 X - Window 工具，也可以作为远程用户使用 GDE 的功能。详细的功能介绍可参考网页http://bimas. dcrt. nih. gov/gde_sw. html 和http://www. tigr. org/~jeisen/GDE/GDE. html。

（3）GeneDoc 这是一种比对的编辑和显示编辑器，由匹兹堡大学超级计算机中心的 K. Nicolas 和 H. Nicolas 编写。GeneDoc 可以引进其他格式，可以通过插入或删除空位来移动残基及拖放编辑。编辑后的比对可以进化树进行重新计算。GeneDoc 可从http://www. psc. edu/biomed/genedoc/获得，在微软视窗上运行。

（4）MACAW 这是集局部比对和编辑于一体的工具。MACAW 从序列中找到无空位板块并计算统计数据。后期的版本允许用户从 3 种方法中选择一种来找到板块。还可寻找用户规定的特殊模式序列。此软件可在视窗、苹果机和其他多种平台上运行。可从http://www. ncbi. nlm. nih. gov/pub/software/shuler/macaw 网页上无记名下载。

除此之外，还有类似如前面提到的 BOXSHADE 等用于格式化的工具，可以用来增强 MSA 的视觉效果。

5.3 多序列比对在生物信息学研究中的应用

5.3.1 序列结构域和基序的寻找

序列基序（motif）指的是一组序列所共有的一段局部保守区域或短的序列模式。其长度范围可以从数个到数百乃至数千核苷酸或氨基酸不等。但绝大多数为较短序列。基序通常可以用来预测分子功能或结构特征或分子亲缘关系。基序可通过多序列比对而产生，可以是固定的序列或是可变的序列。常通过位置特定记分矩阵和隐藏马克模型（HMM）以剖

面的形式存在。与基序不同,结构域是指一组序列中保守的一段较长的序列,绝大多数用于蛋白质序列。可以是蛋白质分子序列的一部分或整个分子。但有些较小的结构域可能和较大的基序交叉。现有的主要与基序和结构域有关的数据库包括:CDD(conserved domain database), Prosite, BLOCKS, PRINTS, PFAM(protein family), COG(cluster of orthologous genes)。其中 CDD,PFAM 和 COG 以结构域为主,而余下的则主要是基序数据库,参见第 6 章有关内容。

5.3.2 基因调节因子预测

随着后基因组时代的到来,我们对分子生物学的研究已经超越对单个基因序列和功能的了解,开始探索基因的调节和从基因组的范畴来理解基因调节的网络。基因表达的主要调节环节是转录的控制。因而理解基因调节,重要的一步是确定每一个基因序列中的调节因子,也就是基因序列中用于与转录因子或其他调节因子结合的区域。已有的研究结果表明,这些转录因子结合区域往往具有一定的序列特征或称基序,而且具有一定的保守性。传统用于确定转录因子结合区域的方法是通过烦琐的重组 DNA 技术辅之报告基因在体外或体内进行功能评估来确定这些因子所在区域的范围。但很显然,用此法来确定 3 万多人类基因的调节因子序列是不现实的。因此,我们急需一种高特异性和高分辨率的方法来确定这些通常是较小的功能性元件(functional element)。由对多种脊椎动物间的序列比较分析已经开始显示出这一方法的可行性。Thomas[10]等在最近的研究中通过对 12 种脊椎动物对应基因组序列的比较分析发现了新的非编码保守区。这些区域未能被实验的方法和成对比对的方法所发现。在这一研究中,他们用一种称为进化遗传印记(phylogenetic fingerprinting)的计算方法来找到很小的从所比较的序列中高度保守的基因组区域。这些高度保守的区域因已经历长时间进化的选择而很可能具有重要的功能。由于相近物种间的序列面临以背景杂音信号为主的问题,在分析中包括亲缘关系较远的多个物种的对应序列应有效地降低杂音。尽管到写稿时能用于种间基因组比较的物种仅限于人、小鼠和大鼠,但根据现已列入基因组计划的物种数目,预计很快能扩展到近 10 种之多。这种对多物种的序列比较分析将为系统地在基因组的范畴上确定基因的调节功能单位提供一个强有力的研究手段。图 5-11举例说明基因上游区域的序列比对所发现的保守区与实验所证实的包含 CRE 和 E-box等在内的增强子元件簇相对应[11]。图中序列代表通过对来自人、小鼠和大鼠肾素基因上游启动子区进行 MSA 比对所获得的保守区之一(作者未发表数据)。图中阴影区域为实验证实存在于人类肾素基因的包含多个启动子的区域。从本图中 MSA 所见,此启动子区域在此 3 个物种中度保守。除此之外,此 MSA 还提示,在此保守区相邻的上游(方框区)还存在至少一个额外的调节因子区可供今后的实验证实。

5.3.3 基因组组装

由于全基因组"鸟枪法"在基因组测序中的应用,在序列的组装过程中,有赖于使用 MSA 从许多的随机 DNA 片断序列中找到相互重叠的序列而逐渐拼接出较大的片断。与前面所述 MSA 的常规应用不同,基因组组装中的 MSA 有以下特点:①重叠区域所涉及的序列理论上属于相同序列,但由于测序错误有可能造成插入或删除。然而正确的序列只有一种,

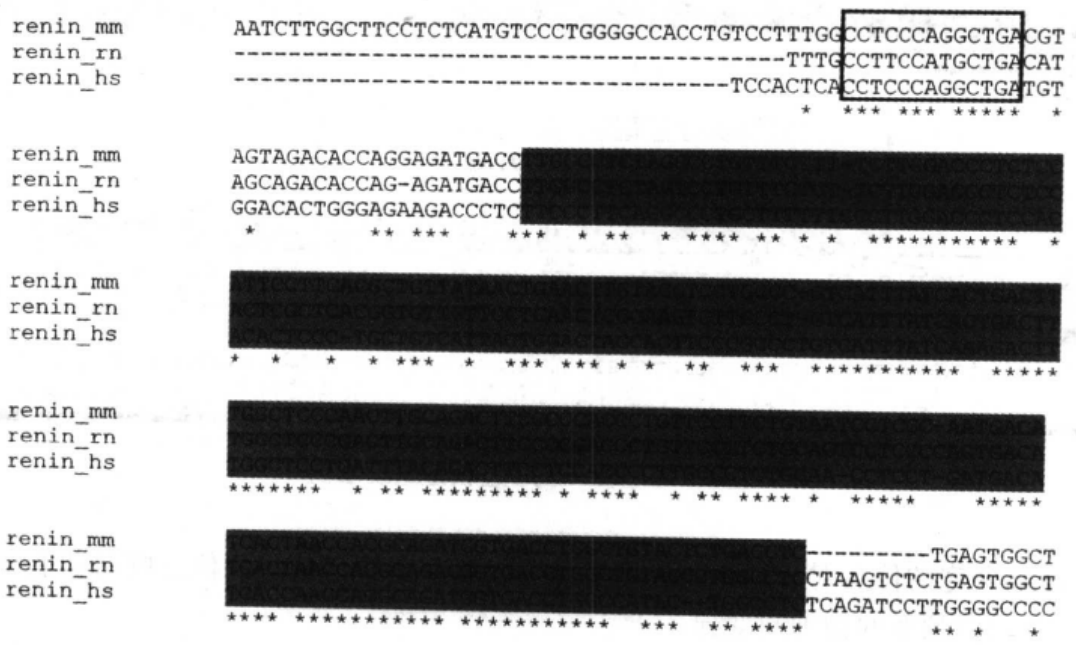


图 5-11 肾素启动子区 MSA 分析结果

renin_hs:人类肾素基因;renin_mm:小鼠肾素基因;renin_rn:大鼠肾素基因

而不像一般的多序列比对存在多种可能性。②所涉及的序列有可能是正向或反向,因而包括对互补序列的比对。③序列间的关系经常是长度未知的重叠或是较大的序列包含较小的序列。常用的商业序列分析软件包通常包括这类功能。如 Vector NTI 中的 Contig Express 就是以此为功能。作为基因组数据的用户,如果想要利用基因组过程的早期数据,如 trace 序列,则需要使用这类工具来获得特定基因的部分甚至全部序列。这是充分利用基因组数据的有效途径,有时可以让用户提前数年获得所需数据,因而具有明显的优势。图 5-12 为基因组组装示范图。本图演示 VECTOR NTI 软件包中的 Contig Express 工具。右上框标示本组装包含 4 个原始序列片断(trace),其名称和序列方向均有标明。右下框显示对应于上框中竖条所辖区域的详细序列。可见,第 1、2 条序列来自同一区域而与第 3 条序列部分重叠,而第 3 条序列又与第 4 条序列在另一端部分重叠,进而形成一个较长的组装层叠组(contig),其推测的共同序列显示于底线。注意:重叠区域的序列在第 1、2 条和第 3 条之间有差异(由箭头所标记的位点)。在此种情形下,因为此区域属第 3 条序列的最末端,为序列错误率较高区域,而另两条序列在此区域完全一致。因此,应当以第 1、2 条的序列为准。

5.3.4 系统发生遗传学分析

生物学中的问题之一是研究生命的起源和现今自然界所存物种之间在系统发生过程中的亲缘关系或称为系统发生遗传学(phylogenetics)或进化遗传学(参见第 8 章)。早期的系统发生遗传学研究大多依赖于形态学和解剖学的数据,但自从生物学研究进入分子生物学时代以来,从分子水平上(DNA 和蛋白质)来研究系统发生遗传学便成为主流。而其中最常见的前期步骤主要是多序列比对。如图 5-13 所示,MSA 中序列间的变化可以用来推测其所代表的物种间的亲缘关系。序列中有些位点的残基对蛋白质的功能非常重要而很少产生

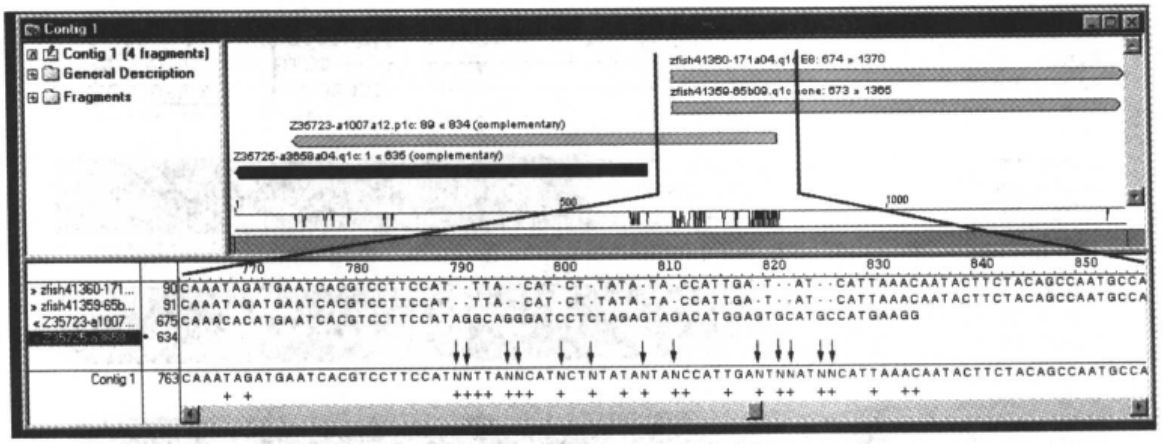


图 5-12　MSA 用于基因组组装示例

突变。这些保守的位点对于产生 MSA 起关键作用。相反,其他一些位点则对功能不那么重要,因而有较高频率的物种间变异。

A　PN-FLSC
B　PN-F-SC
C　PNKYLSC
D　PN-YLSC

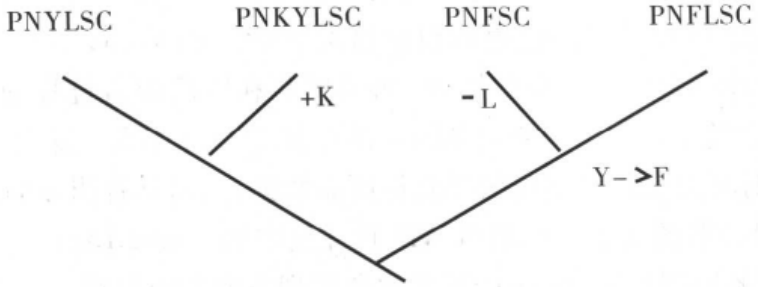


图 5-13　MSA 用于系统发生遗传学分析

系统发生遗传学研究中要考虑的问题包括大多数物种在起源上的复杂性。一般情况下,基因组的一部分基因由正常的生殖周期进行垂直遗传,而另外的部分可能由包括病毒、DNA 转化、共生等多种机制导致的基因水平转移(Horizontal gene transfer)而来。使用水平转移的基因序列会扰乱系统发生遗传学的分析。基于以上原因,用于系统发生遗传学的基因最好具备以下特点:基因普遍存在于绝大多数物种而且容易通过其序列的保守性被识别。与此同时,这些基因序列应当有足够的变异来区分亲缘相近的物种。在已知的 DNA 序列中有两类分子符合上述条件,携带大量的种间变异而被广泛用于系统发生遗传学的研究:核糖体 RNA(rRNA)小单体和线粒体序列。现有的普遍公认的细菌进化树大多是根据 16SRNA 序列而建立,而线粒体序列较多被用来研究灵长类的进化。

除研究物种间的亲缘关系之外,MSA 和由其衍生的关系树可以用来探讨同一基因组中同系基因(paralogous genes)间的进化关系或较大蛋白家族内的分类[12,13]。在此情形下,使

用的序列通常为蛋白质序列。这类分析可以用于预测蛋白的功能和进化关系。

（梁 平）

参 考 文 献

1 Dayhoff MO. Atlas of Protein Sequence and Structure. Washington: National Biomedical Research Foundation, 1978. 345 ~ 352

2 Henikoff S , Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA, 1992, 89: 10 915

3 Needleman SB, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 1970, 48: 443

4 Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol, 1981, 147: 195

5 Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 1988, 73: 237

6 Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position – specific gap penalties and weight matrix choice. Nucl Acids Res, 1994, 22: 4 673

7 Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins, 1990, 7: 41

8 Krogh A, Brown M, Mian IS, *et al*. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol, 1994, 235: 1 501

9 Eddy SR. Multiple alignment using hidden Markov models. Proc Int Conf Intell Syst Mol Biol, 1995, 3: 114

10 Thomas JW, Touchman JW, Blakesley RW, *et al*. Comparative analyses of multi – species sequences from targeted genomic regions. Nature, 2003, 424: 788

11 Pan L, Xie Y, Black TA, *et al*. An Abd – B class HOX. PBX recognition sequence is required for expression from the mouse Ren – 1c gene. J Biol Chem, 2001, 276: 32 489

12 Liang P, Labedan B, Riley M. Physiological genomics of Escherichia coli protein families. Physiol Genomics, 2002, 9: 15

13 Liang P, Riley M. A comparative genomics approach for studying ancestral proteins and evolution. Adv Appl Microbiol, 2001, 50: 39

第6章　蛋白质结构分析与结构生物信息学

本章是在第5章详细介绍蛋白质序列分析的基础上，进一步介绍蛋白质结构分析方法。第2章中，我们已经知道，蛋白质结构可分为一级、二级、三级和四级结构，这是蛋白质结构分析的基础（详细内容希望读者参考其他生物化学教材中有关蛋白质结构的论述）。在已知的大量蛋白质实验数据的基础上，通过生物信息学方法建立相应的蛋白质结构数据库，并进一步比较和分析其结构特点，归纳总结出结构模型，编写出相应程序，在反复验证的基础上，用于新的蛋白质的结构预测，这便是目前生物信息学最具挑战的领域之一，随之诞生出生物信息学的重要分支——结构生物信息学（structural bioinformatics）。

结构生物信息学是在原子和亚细胞空间水平重点讨论生物大分子结构信息的表达、储存、检索、分析和显示方法。结构生物信息学的目标有二：一是建立生物大分子信息处理的通用方法，二是利用这些方法解决生物学中的问题和得到新的预测性知识。由于近年来高通量分析技术的应用和大量生物大分子结构数据的出现，大大加快了结构生物信息学的发展，也使得生物学在更高的层面进行分析，如结构模型的建立、蛋白质新功能的预测、药物靶向设计、分子模拟等。正是基于这些结构分析的基础，许多生物现象可以进行直视（visualization）、分类（classification）、预测（prediction）和模拟（simulation）。由于分子结构所得到的数据往往是非线性的，其空间变量又是连续性的，加之相互作用分析中需引入许多重要的物理学原理，以及3D模型建立中的图形处理与分子模拟的复杂性等，使得结构生物信息学较传统生物信息学中所用方法和模型更具挑战性[1]。本章将结合蛋白质结构分析常用方法阐述结构生物信息学的基本内容及其应用。

6.1　蛋白质基本特性分析

通常在获得蛋白质一级结构后，可对其基本特征（参数）进行分

析，包括分子质量、分子式、等电点、氨基酸组成、理论消光系数、疏水性、稳定性等物理、化学特征。最常用的分析软件是 ProtParam（http://us.expasy.org/tools/protparam.html）。读者可将自己感兴趣的蛋白质序列输入该程序，得到该蛋白质的基本特性。这些特性对于相关实验具有重要的指导意义，尤其是二维电泳中具有重要参考价值。预测所得到的分子质量与实验结果可能有所出入，主要由于翻译后修饰的结果。

6.2 蛋白质结构域、基序与结合部位分析[1~4]

蛋白质结构域（domain）是结构生物学、遗传学、生物化学和进化生物学中的重要概念。在不同学科中，其定义常有所不同。结构生物学中结构域被定义为多肽链上那些可折叠为球状结构的特殊片段，并发挥特殊的生物学功能。但在遗传学上结构域常用来表示基因中发挥某种功能的最小片段。基序（motif）通常是指蛋白质相互作用中最小的功能单位，一般位于球形蛋白质的表面。结合部位（binding site）则是指蛋白与蛋白或其配体结合的具体部位，通常是一个或几个氨基酸残基。不同蛋白质之间相同或相近的结构域被称做模块（module）。据估计，人类基因组中含 1 000~6 000 个这种模块。模块可独立存在于一些较小的蛋白质之中，也可与其他模块一起存在于多肽链上，这通常是由于基因的融合所引起的。一组模块常共同表达，存在于同一细胞或亚细胞器中。模块间的相互作用是蛋白-蛋白相互作用的基础。后基因组时代的任务之一是在系统生物学水平解析这些模块及其相互作用，从而更好地理解蛋白质的生物学功能。

蛋白质结构生物信息学的任务之一便是通过已有的蛋白质 3D 实验数据和其他已知的功能信息，建立结构域的数据库和相应预测模型。最初阶段，结构域的鉴定是通过直观和作图分析拆分出不同结构域的界限。其出发点是基于这样一个事实：结构域内的原子间接触要比结构域间的原子接触更紧密。当然，完全套用这个模式也会带来一些误差。结构域分析的算法研究自 20 世纪 70 年代开始，经历了碳原子间的共价结合、侧链接触等物理、化学模型和聚类分析、主成分分析等统计模型，到理论图形方法等不同的发展阶段。要统一一种标准用于结构域分析是很困难的。结构域分析的主要数据来源于 SCOP 和 CATH 等结构分类数据库（参见 6.8.2 节）。通过 PredictProtein（http://cubic.bioc.columbia.edu/predicprotein）服务器可获得有关结构域、基序和结合部位的大部分信息。

6.3 蛋白质拓扑结构、折叠和三维结构模型[1~4]

当观察一个蛋白质的 3D 模型时，总会发现一些重复出现的有规律的结构，如螺旋或线性结构。由 3D 模型来推断这种次级结构特征，并建立相应的模型来预测这种次级结构特征是目前结构生物信息学的热点之一。这些次级结构特征的主要用途有：① 预示折叠方式；②蛋白质结构视观中的直觉方式；③影响序列的比对；④ 与功能密切相关。利用生物信息学方法建立一个好的次级结构分析程序，既能与已知结构保持一致，又能较准确地预测未

知蛋白质的结构和功能，这是结构生物信息学的重要挑战领域之一。目前已建立的程序中，DSSP[5]（http://www.cmbi.kun.nl/gv/dssp）是通过3D模型推断和预测次级结构最好的程序之一，它是基于骨架之间的氢键模型建立的程序，这也是通过解析α-螺旋和β-片层（参见图6-1）不同状态间形成氢键的方式而推断。

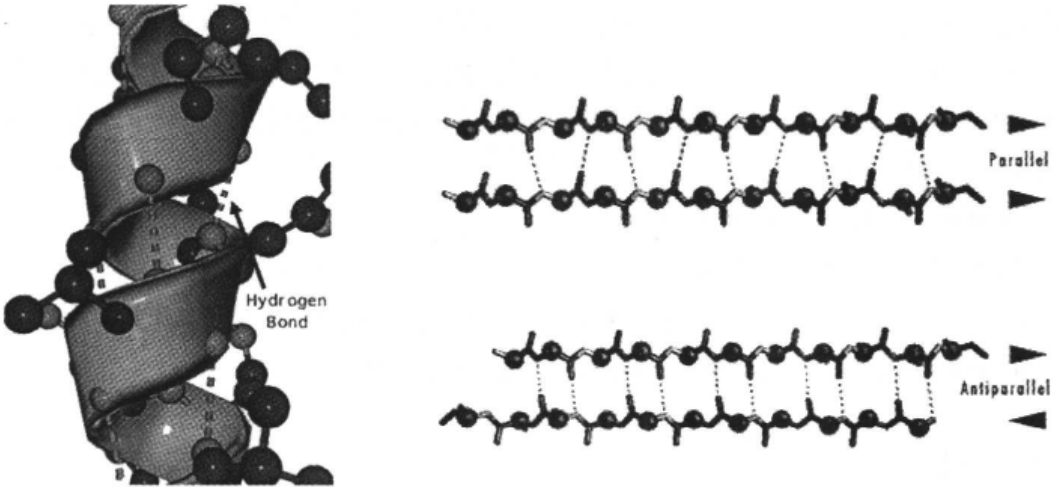


图6-1　蛋白质二级结构中的α-螺旋和β-片层结构示意图

由蛋白质1D预测3D是结构生物信息学工作者面临的一大挑战，是CASP（Critical Assessment of Structure Prediction）总结有关结构预测方法领域中取得的重大进展。从1994年以来，CASP每隔2年更新一次，到2002年，已运作到CASP5。CASP主要针对蛋白质折叠预测方法的评估而建立。总体而言，现有的关于蛋白质拓扑结构、折叠方式和3D结构的预测，其准确性约70%左右。

蛋白质的折叠预测方法主要分为3类：同源性模型（homology modeling）、折叠识别（fold recognition）和从头开始折叠（*ab initio* folding）。同源性模型是根据同源蛋白质的结构分析得到有关结构域和相应的结构特征，再预测其折叠方式，是目前蛋白质3D结构预测中最常用的方法。折叠识别又称反向蛋白折叠（reverse protein folding），将待分析蛋白质序列作为查询单位，查找几种常用的折叠数据库，如SCOP、CATH和DALI等（参见6.8.2节），再对其3D结构进行预测。从头开始折叠是直接根据蛋白质序列的理化特征预测其构象的方法。

6.3.1　同源性模型

从实验所得到的蛋白质模型结构，经计算机处理并存储于数据库，便于研究者查询和比较，这无疑是进一步研究该蛋白质功能和进行药物靶向设计的理想工具。事实上，就人类基因组而言，新基因尚在不断发现之中，其结构尚未测定。即使对于那些已发现的基因，特别是当它们编码一些大分子质量蛋白质的时候，现有的技术如晶体结构分析和NMR技术，尚难以对其结构进行测定。这时主要依靠结构生物信息学方法对其3D结构进行预测。同源性模型是目前常用的3种3D结构预测方法中最简单也最常用的方法，是结构生物信息学工作者必须掌握的基本方法。其基本原理有两点：①结构是由序列所决定；②进化过程中，结构上的变化相对序列变化更缓慢。也就是说，相似的序列往往会得到相似的结构，进化上

距离较远的序列也可能折叠成相似的结构。

假定我们想知道一个含150个氨基酸的蛋白质(A)结构,我们首先将该序列与PDB数据库中的已知序列进行比较(例如用BLAST)。幸运的话,我们发现待分析序列(A)与PDB中的一个含300个氨基酸的蛋白质(B)结构具有约50%的相同性。这时从PDB中调出B序列,将同源区切割出来,并将不相同的氨基酸残基位置进行突变替代,这样便得到A序列的预测模型。总结一下,可将同源性模型的分析过程归纳为以下步骤:① 模板识别和比对分析;② 比对校正;③ 骨架产生;④ 环状模型;⑤ 侧链模型;⑥ 模型优化;⑦ 模型确认。在以上分析中,常用的程序分别有:

(1) FASTA 或 BLAST(详见第4章);

(2) CLUSTALW(详见第5章);

(3) SWISS-MODEL(http://www.expasy.ch/swissmod/SWISS-MODEL.html);

(4) MODBASE(http://guitar.rockefeller.edu/modbase)。

6.3.2 折叠识别

折叠识别分析原理是根据识别出的已知蛋白质的折叠区,来对未知结构蛋白质的折叠结构进行指认,即寻找适合某些折叠结构的序列(反向蛋白折叠),或称牵线(threading)。折叠识别方法中最重要的一点是,表面看来不相关的蛋白质可能具有相似的折叠方式。其原因可能是:① 趋异进化(divergent evolution),即本来相关的蛋白质,进化过程中序列变化较大,但其折叠方式变化不大;② 趋同进化(convergent evolution),即由于功能上的需要,功能相同或相近的蛋白质其结构具有相似性;③ 序列变化千千万万,但折叠方式屈指可数,导致不相关的蛋白质,其折叠结构也具有某种相似性;④ 折叠分类方法的错误,产生结构的相似性,即得到所谓假阳性结果。现有的折叠数据库如SCOP、CATH、DALI和FSSP等(参见6.8.2节)采用不同的聚类方法。这些数据库是折叠识别的基础。折叠识别中先对待分析序列查询这些数据库,找到相关折叠方式和序列,再将待分析序列的结构拼接。折叠识别常用的程序或方法包括PSI-FASTA、HMM、SAM、THREADER、3D-1D、123D等。

6.3.3 *ab initio* 折叠

ab initio 折叠即完全依据蛋白质的氨基酸序列预测其构象。其分析原理是基于蛋白质一级结构决定其空间结构。这是物理化学的重要分支。成功的预测需要利用自由能函数,并满足最小化原理,即天然状态下自由能最低,所形成的构象空间所需要的自由能也应该是最低。这种分析方法的重大挑战在于建立较精确的预测模型。多肽链画像(representation)、势能函数(potential function)、粗略搜索(coarse searching)等是常用的预测方法。这些方法也被用于基因组的组装分析和结构基因组分析。

6.4 蛋白质结构视觉和分子模拟

6.4.1 蛋白质结构视觉

蛋白质结构视觉(visualization)是在实验测定其结构或通过结构生物信息学进行结构预测的基础上,对蛋白质结构利用计算机图形处理方法显现出来,便于研究人员对其二维或三维结构有一感性认识,更重要的是有助于理解蛋白与蛋白或其配体的相互作用。因此蛋白质结构视觉和分子模拟是研究蛋白质复合体、代谢途径、酶动力学和药物设计等的重要工具。

蛋白质结构视觉形式主要分为原子代表模型(atom representation)、摘要观察(abstract view)和表面观察(surface view)3种,分别以RasMol,MolScript和AVS为代表。最早的结构视觉出现于20世纪60年代末,伴随计算机图形处理软件和程序的发展,视觉形式和效果有了很大的改进,并逐渐引人动态视觉效果。目前已出现许多相应视觉软件和网络服务器。基本上是根据结构数据库中已有的模型,编译相应的视觉程序。大致分为3类:模型建立、模型查询和结构呈现。结构呈现软件中最流行的是RasMol(http://www.umass.edu/microbio/rasmol/),已与PDB链接。表6-1所列为目前常用的蛋白质视觉软件[1]。

表6-1 目前常用的蛋白质视觉软件

软件名称	功能	评价	优点	缺点	操作平台
AVS	结构查询和呈现	表面观察最常用的软件	非常灵活,具有多个模块,可进行多种数据分析	价格昂贵,操作复杂,不易掌握	Linux, SGI, Unix
Chimera	用于基本的模型建立、查询和呈现	功能齐全的视觉软件	非常灵活,功能强大,适合多种数据资料	源程序封闭,尚在完善之中	NT4.0, W2000, Linux, SGI, Unix
MICE	用Java语言编译的结构呈现程序	基于网络界面,可进行合作研究	通过ID号可接入到PDB中的所有数据	需要Java3D程序支持,某些操作平台无法运行,查询功能受限	NT4.0, W2000, Linux, SGI
MolScript	结构呈现	显示图形和相互关系。Bobscript为其升级版	程序命令丰富,输出格式多样,可进行预览	程序命令隐晦	Linux, SGI, Unix

续表 6－1

软件名称	功能	评价	优点	缺点	操作平台
O	电子密度图像和模型建立	功能强大的模型建立软件	非常全面的模型建立功能，也具有某些图形呈现功能，具有灵活性	不易掌握	NT4.0，SGI
Quanta	模型建立	工业用户常用的模型建立和修改软件	功能强大	价格昂贵，软件较旧	SGI
QuickPDB	结构查询	用 Java 编译的显示结构和序列的工具	网络浏览，可同时显示序列	非常简单	适合各类计算机
RasMol	结构呈现	功能齐全的结构观察软件，Chime 为其商业版，可在 Netscape 中操作，RasTop 是为 Windows 用户设计的增强版	操作简便，回应快捷，功能全，显示方式多样，源程序开放，可由用户自行维护	一次仅显示一种结构，图形效果差	适合各类计算机
VMD	结构分析和呈现，主要是分子动力学	包括各种分子动力学模型，适合一般用途的交互观察	鼠标控制，使用方便	在一些图形功能较差的计算机上显示很慢	Linux，SGI，Unix
WebMol	用 Java 语言编译的结构呈现程序	快捷的图形显示软件	网络浏览，可从 URL 上载结构	显示方式单一	适合各类计算机

6.4.2 蛋白质分子模拟

通过晶体分析和 NMR 获得的蛋白质结构模型是静态的，但其生物学功能的发挥在结构上经历不同状态的动态变化，这往往涉及不同生理条件下能量的变化和蛋白质折叠形态的变化。通过体内直接观察这种结构的动态变化并记录下来往往很困难。蛋白质分子模拟（simulation）就是通过计算机分析蛋白质结构的静电场来模拟这种动态变化，从而在更高层面了解蛋白质的功能。

6.5 蛋白质相互作用分析

体内蛋白质发挥其生理功能时，通常是与其他蛋白质协同作用。绝大多数蛋白质是与

其他蛋白质一起,形成紧密结合(亚基)或松散型蛋白质复合体,组成某一网络或代谢途径。认识蛋白质的相互作用是解析其功能的基础。分析蛋白质相互作用的方法目前主要依靠遗传学或生物化学方法,如酵母双杂交或三杂交系统、免疫共测定技术、质谱分析等,目前这些方法都建立了相应的高通量方法。随着结构信息学的发展,相互作用的预测也成为相互作用组(interactome)的重要内容。

相互作用的预测主要包括结合部位的预测和结合对象的预测。几种常见的预测方法有:①进化遗传分析,如进化遗传档案、Rosetta - Stone 法、染色体邻近分析等;②从头预测(*ab initio* prediction);③贝叶斯网络法(Bayesian networks)[6]。依据这些分析建立的相互作用数据库包括 SPIN - PP、MIPS、DIP、BIND、ProNet 等。

6.5.1 进化遗传分析

在蛋白质相互作用分析中,通过与进化遗传学分析方法结合,可以得到有关进化保守区的相关信息,特别是结合有关突变分析资料,从而推断某些相互作用的部位。这些进化保守区常属于结构重要区或活性部位。相互作用部位的预测可以依据结构进行物理锚定,也可以依据序列相关性突变分析,或者是结构与序列相结合的杂交方法。相互作用对象的预测分为两大类:① 基于基因组信息。包括种系发生档案(phylogenetic profiling)、邻近基因保守性(conservation of gene neighboring)、基因融合(gene fusion) 或称 Rosetta - Stone 法等 3 种主要类型。参见图 6 - 2。② 基于序列信息。包括相关性突变(correlated mutation, i2h)和种系发生树相似性(similarity of phylogenetic tree, mirrortree)。

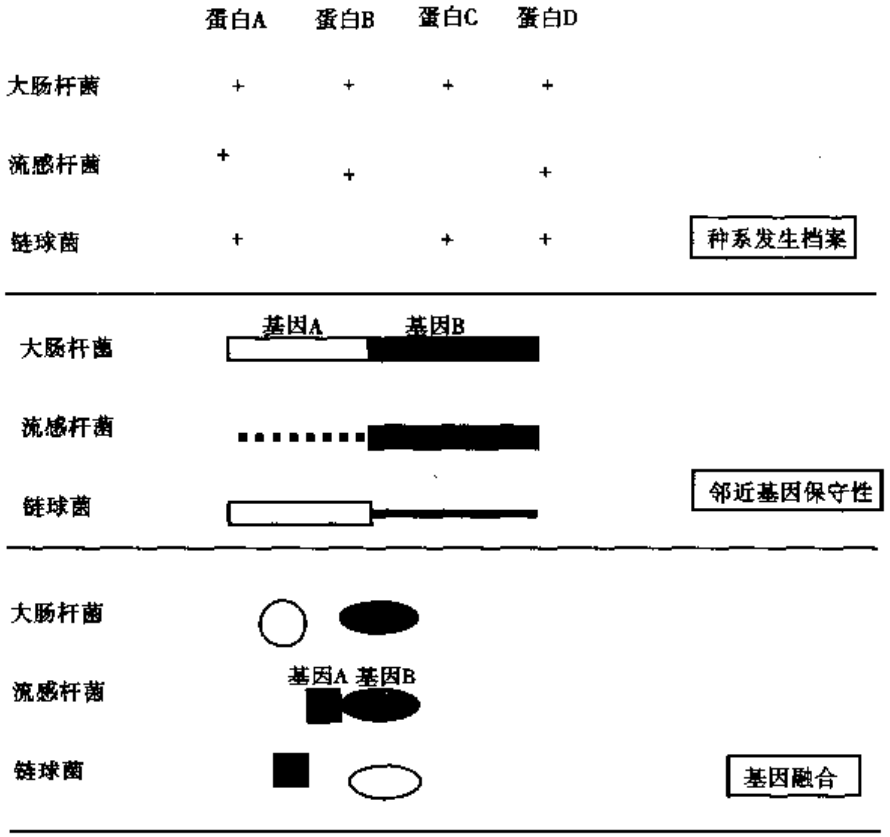


图 6 - 2 基于基因组信息的进化遗传分析蛋白质相互作用

图6-2是3种基于基因组信息的预测方法的示意图。种系发生档案分析中,蛋白A和蛋白C在不同种属中总是同时存在或同时不存在,表明这两种蛋白相互协同作用。邻近基因保守性分析中,在大肠杆菌基因组中可以看出A和B基因在染色体上的邻近关系,也是相互作用的证据。基因融合分析中,在流感杆菌中出现A和B的融合基因,往往是A和B相互作用的证据。

相关性突变研究(i2h)既可用于相互作用部位的预测,也可用于相互作用对象的预测,只是分析序列的大小不同而已。通过多序列队列分析,针对筛选出的两个序列,分别比较3种类型的配对关系:AA、AB、BB,得到相互作用指数,根据其相关性判断A和B是否存在相互作用。mirrortree分析是比较A和B两种可能相互作用的蛋白的进化树。如果两者存在相互作用,其进行过程应该相似,即共进化(coevolution)。根据各树枝所得到的进化距离,列成阵列,然后按照遗传距离作图,观察A和B之间遗传距离的相关性,来说明是否存在相互作用。以上两种基于序列分析的进化遗传分析方法可以看做种系发生档案分析的深入,即从主干分析深入到树枝的分析。

6.5.2 *ab initio* 预测

蛋白质相互作用的从头预测是根据多肽链的物理化学特征预测其折叠方式,再通过所得到的模型,预测其相互作用区域或相互作用对象。要建立一个较准确的从头开始预测程序是结构生物信息学的一大挑战,参见6.3.3节。静电相互作用分析(electrostatic interaction)可视做从头开始预测方法之一,这也是分子能量学的重要内容,特别是在配体锚定(ligand docking)分析、蛋白质复合体预测和药物设计中有重要应用价值。静电相互作用分析中最重要的分析原理是利用线性Poisson-Boltzmann方程(PBE),这同时也考虑溶液环境的影响。已有一些程序如DelPhi、AMBER等就是利用PBE进行静电相互作用分析的。

6.5.3 贝叶斯网络法

贝叶斯网络法[6]是最近建立的分析酵母蛋白质相互作用的方法。该法利用MIPS数据库中蛋白质复合体作为阳性对照,同时利用定位于不同细胞器上的蛋白质作为阴性对照,建立训练和测试模型。然后整合有关高通量酵母双杂交系统和免疫共沉淀的实验资料和有关基因组水平相互作用的资料,利用贝叶斯定理建立相互作用的概率预测模型,并经相关实验认证。这是尝试将不同数据进行整合,以提高预测准确性的一种方法。不同方法的整合将为相互作用组提供高丰富、更准确的内容。图6-3总结了目前蛋白质相互作用分析的不同方法。

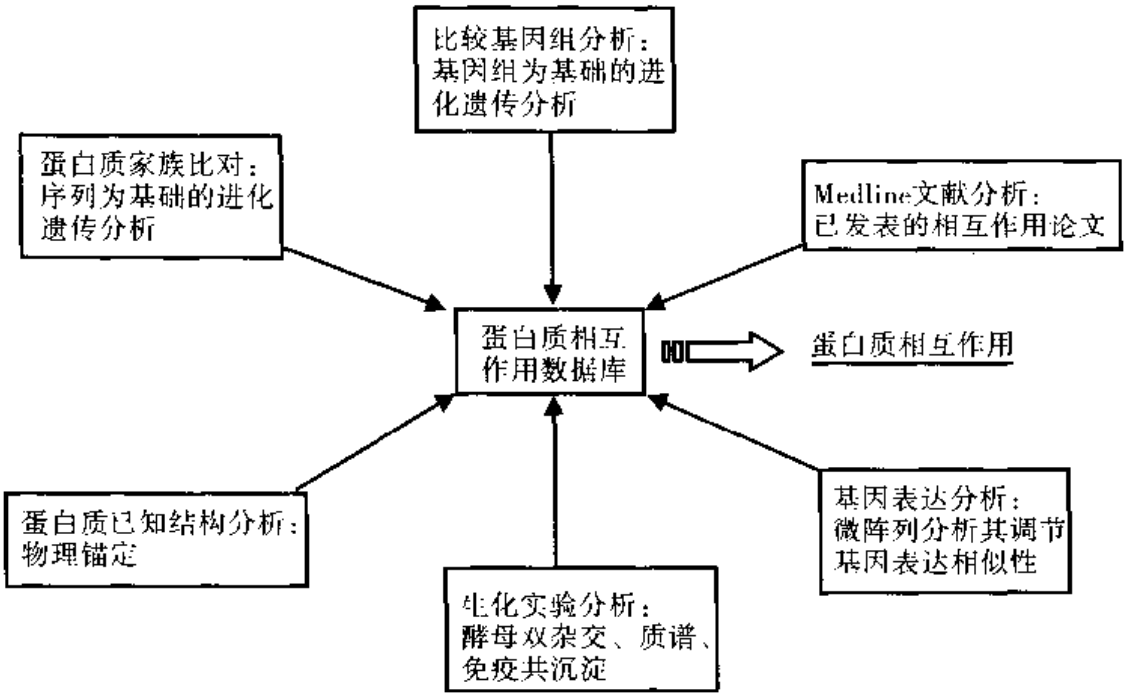


图6－3　蛋白质相互作用分析方法

6.6　蛋白质结构分析的基本流程

结构生物信息学中涉及的软件和数据库种类繁多，如何从中遴选出比较实用、结果较可靠的分析方法是用户最实际的问题。特别是针对一种新蛋白质，如何通过预测获得有用的结构信息，以指导实验设计。优化一套蛋白质结构分析的流程往往可起到事半功倍的效果。这里笔者结合自己对 HPS 蛋白分析的体会，总结以下分析流程，供读者参考，见表6－2。蛋白质分析中的趋势是将各种程序整合到服务器上，可得到多种不同的信息。常用的服务器有：

（1）ExPASy：http://us.expasy.org；

（2）PredicProtein：http://cubic.bioc.columbia.edu/predicprotein；

（3）PIR：http://pir.georgetown.edu。

表 6-2 新发现蛋白质结构与功能预测流程表

分析内容	分析软件或服务器	网址	评价
基本特征	ProtParam	http://us.expasy.org/tools/protparam.html	可得到蛋白质一级结构的基本信息
信号肽	SignalP	http://www.cbs.dtu.dk/services/SignalP	预测蛋白质的信号肽
结合部位	ScanProsite	http://us.expasy.org/tools/scanprosite	检索 Prosite 数据库的快捷方法
结构域	SMART	http://smart.embl-heidelberg.de	结构域的常用数据库
	InterPro	http://www.ebi.ac.uk/interpro/scan.html	检索 InterPro 数据库的方法
拓扑结构	MATA	http://cubic.bioc.columbia.edu/predictprotein/submit_met.html	该服务器可自动链接到不同的拓扑结构分析程序，如 PROFphd、PSIPRED、SSpro2 等，通过 E-mail 反馈结果
跨膜结构	HMMTOP	http://www.enzim.hu/hmmtop	常用的跨膜结构预测程序
	TMHMM	http://www.cbs.dtu.dk/services/TMHMM-2.0	常用的跨膜结构预测程序
	TMPred	http://www.ch.embnet.org/software/TMPRED_form.html	常用的跨膜结构预测程序
细胞定位	PSORT	http://psort.nibb.ac.jp/form2.html	查找细胞定位信号或基序
线圈结构	COILS	http://www.ch.embnet.org/software/COILS_form.html	主要预测线圈绕线圈结构
3D 结构	PDB	http://www.pdb.org	新发现蛋白质通常为阴性结构，但可与同源蛋白质进行结构比较
	CE	http://cl.sdsc.edu/ce.html	3D 结构的比较
相互作用	MIPS	http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction	主要收集酵母中蛋白相互作用
	DIP	http://www.ampere.doe-mbi.ucla.edu:8080/dip.html	包括各种不同类型的蛋白相互作用，并与原始文献链接
同源性与蛋白家族	BLASTp/Swiss-Prot	http://www.ncbi.nlm.nih.gov/BLAST	通过 BLAST 查找同源蛋白及结构域

6.7 蛋白质结构分析的技术平台

目前蛋白质结构分析主要有以下几大技术平台：结构生物信息学、X－衍射蛋白质晶体结构分析、核磁共振(NMR)光谱分析、电镜技术。此外，高通量的酵母双杂交系统、免疫共沉淀技术和质谱分析方法可视作生物化学中蛋白质相互作用分析的主要技术平台，也是结构基因组学的有力工具，可参考相关实验手册，此处不再赘述。

6.7.1 X－衍射蛋白质晶体结构分析[1]

X－射线晶体结构分析(X－ray diffraction crystallography)是解析生物大分子结构与功能的基本方法。该法首先是将待分析的纯化蛋白质形成晶体，然后利用X－衍射技术得到该晶体的相关数据，整合成相应的图像，存储于结构数据库中。该法耗时费力，特别是需要大量人工处理各种数据，和将来源于不同软件的资料加以整合。这样，快速处理晶体结构数据和减少差错便成为目前蛋白质晶体结构分析的两大难题或瓶颈。近年来，随着自动化和高通量分析技术的应用，大大加快了蛋白质晶体结构分析。例如，剑桥结构数据库(CCSD)每年新增大约15 000条晶体结构数据条目。同时，传统的X－衍射蛋白质晶体结构分析技术也得以大大改进，提高了分析效率。尽管如此，晶体数据处理中的运算法则仍需要完善，数据的加工处理和分析涉及大量统计生物信息学的工作。晶体结构分析的常用软件有SOLVE，RESOLVE等。另外，该技术仍无法解决某些难以形成晶体结构的蛋白质的分析。

高通量晶体结构分析中的几大重要环节是：数据处理与分析、重原子的定位、密度修饰、分子替换、图形整合、模型加工和确认。晶体结构分析中的自动化所面临的主要挑战是非晶体结构的对称(noncrystallographic symmetry，NCS)和紊乱现象(disorder)。随着近年来晶体结构的运算法则和计算机科学的发展，相信新一代的自动化分析软件将进一步解决以上所面临的问题，并将适时处理各种衍射数据和加快图形整合过程。

6.7.2 核磁共振光谱分析[1]

大分子结构分析的核磁共振光谱分析(nuclear magnetic resonance (NMR) spectroscopy)是结构基因组的重要内容，约占PDB中15%左右。特别在测定蛋白质与配体结合的相互作用时，NMR技术起重要作用。同时，NMR也是酶动力学研究、蛋白质折叠和稳定性研究中的重要工具。随着NMR的有效性和应用范围的拓展，其在结构基因组或结构生物信息学中的应用也愈加广泛。目前美国NIH属下的NIGMS研究所成立的蛋白质结构协作组(PSI)中，NMR成为一些中心的支柱性技术。

与X－衍射晶体分析技术相比较，NMR技术无须制备晶体标本，可在溶液中直接测定，也可进行固相测定，因此利用NMR法使得某些无法获得晶体结构的蛋白质或非液相蛋白质(如膜蛋白)的结构测定成为可能。相对而言，NMR技术更适合小分子质量蛋白质结构的分析，目前主要来源于大肠杆菌中表达质粒所表达的纯化蛋白质。对于蛋白质折叠、局部动力学或构象分析、蛋白－蛋白相互作用，NMR更体现其优越性。NMR的不足之处是成本太

高，步骤繁多。因此，两者在结构基因组中具有互补性，常有机地结合在一起。一般分子质量大于 60 ku 者用 X－衍射晶体分析法，分子质量小于 60 ku 者用 NMR 法。

NMR 光谱技术的原理是分析受磁场作用下，经磁力加速旋转的原子核不同状态间转换时的情形。常用于蛋白质 NMR 研究的原子核包括^{1}H，^{13}C，^{15}N 和^{31}P。结构测定中，首先是利用多维、多原子核 NMR 方法测定序列特异性结构信号。这种结构信号是二级结构的基本特征，也能反映出半胱氨酸的氧化状态，并用于测试或确认所得到的结构模型。其次，通过多种不同的 NMR 实验获得其他结构区域的信息，如原子核间的距离、转角、排列方向、氢键形成等。最后将所有 NMR 资料加工处理，得到的数据存储于 PDB 中。

PDB 中的有关蛋白质 NMR 条目包括：所研究蛋白系统的描述、蛋白所处的溶液环境（溶剂、pH、温度、压力、浓度等）、NMR 数据采集方法、NMR 参数和相关信息以及未加工的原始资料等。

NMR 的数据处理涉及许多生物信息学方法。首先是将原始资料进行 Rourier 转换为不同的峰值，然后采集各种不同的峰组成图谱，并筛选出具有特定结构特征的图谱。这些过程常用 NMRPipe 和 SPARKY 软件（http://www.cgl.ucsf.edu/home/sparky/）处理，也使用 XEASY、DYANA 和 GARANT 等软件分析侧链或骨架结构。最后对于所得到的结果或模型需要进一步修正和确认，才能送交到 PDB 中。NMR 技术的发展趋势是提高自动化水平，减少数据处理时间和差错，利用无细胞系统以减少样品处理时间。

6.7.3 冷冻电镜技术[1]

上文述及的 X 射线晶体分析和 NMR 光谱技术主要针对单个蛋白质结构的解析。事实上，生物系统是一个非常复杂而有机结合的系统。生物大分子的装配过程可以提高电镜技术进行观察，并追踪其动力学过程。这种基于电镜技术的结构分析可在接近生理条件下同时观察多种成分的图像变化。这主要是依赖上世纪 80 年代以来的冷冻电镜技术（electron cryomicroscopy）来观察生物大分子的空间构象，同时需要借助生物信息学方法如模式识别（pattern recognition）、数据库分析、同源性模型（homology modeling）等。由冷冻电镜技术所获得的蛋白质三维结构与 X 射线晶体技术非常相似，而且其信噪比非常低，并适合于内在膜蛋白的分析。此技术目前应用面并不太广，也没有形成相应的数据库，尚在不断发展之中，并试图与 X 射线晶体分析和 NMR 技术相结合，以弥补其分辨率的不足。

6.8 常用蛋白质结构数据库介绍

第 4 章已对常用分子生物学数据库的结构和基本类型作了详细论述，这里主要介绍几种常用的结构数据库，作为补充。主要有存储蛋白质结构的 PDB（Protein Data Bank），进行蛋白质结构比较的 SCOP 和 CATH，以及存储次级结构为主的 Pfam 等。

6.8.1 PDB

PDB（http://www.pdb.org）始建于 1971 年，由 Walter Hamilton 在 Brookhaven 国立实验

室创建,用于满足生物大分子结构存储的需要。1998 年 10 月,PDB 由结构生物信息学研究协作组(RCSB)统一管理。其独特的数据录入方式已被称为 PDB 格式,包括原子坐标、化学和生化特征、结构测定实验的详细条件、结构特征(如二级结构、氢键、活性部位、分子装配等)。1996 年 Callaway 等[7]将 PDB 的格式进行了统一修订和描述,成为目前 PDB 数据库的统一格式。PDB 也兼容晶体分析数据格式(mmCIF),目前正在建立相应的 NMR 数据格式。PDB 的备注(REMARK)栏中特别注明该条目的变化和修改情况。

PDB 主要收录 NMR、晶体分析、冷冻电镜和模型预测所得到的结构数据,目前许多杂志要求注明相关的 PDBid 号。PDB 迄今已收录 1 万 7 千多个条目,平均每周增加约 50 条,预计到 2005 年,总数将达到 3 万 5 千条。PDB 的数据处理包括数据存放、分析和确认等过程。数据存放后会自动指定一个 PDBid 号,经过几步确认该数据后,才成为真正的 PDB 条目,其过程约 2 周左右。

PDB 的检索可查询单一结构或多个结构。查询方式分 3 种:①状态查询(Status Query):主要查询所存放数据的处理状态;②SearchLite:用于关键词查找;③SearchFields:属于更高级的查询方式,包括 PDBid、引文作者、序列、实验方法、EC 编号等。查询结果可直接显示或下载。

与 PDB 密切相关的 3 个数据库是生物大分子晶体数据库 BMCD (Biological Macromolecule Crystallization Database, http://www.bmcd.nist.gov:8080/bmcd/bmcd/bmcd.htm),主要存放 DNA 和 RNA 结构的 NDB (Nucleic Acid Database, http://ndbserver.rutgers.edu),以及和三维结构比对的 CE (Combinatorial Extension of the Optimum Path, http://cl.sdse.edu/ce.html)。另外,NCBI 也建立了一个可链接参考文献、序列和基因组信息的结构数据库 MMDB(MacroMolecular DataBase, http://www.ncbi.nim.nih.gov/Structure),该数据库主要源自 PDB,可通过 Entrez 检索。

6.8.2 SCOP 和 CATH

6.8.2.1 SCOP

蛋白质的结构反映其功能和进化史。其结构特征往往通过与其他蛋白质的比较而获得。几乎所有蛋白质都与其他蛋白质具有结构的相似性,或者说具有相同的进化起源。这些相互关系的研究和结构特征的获得是结构生物信息学和分子进化学的重要内容。基于这种需要,Murzin 等[8]于 1995 年建立了蛋白质结构分类数据库 SCOP (Structural Classification of Proteins, http://scop.mrc-lmb.cam.ac.uk/scop)。该数据库主要依据决定三维结构的特征区域的进化学相关性。其基本理念是,不同蛋白质序列分析即使进化关系较远(或者序列相似性很低),但其结构上的构造关系和拓扑结构相对更为保守。SCOP 中的每一条目包含坐标、结构图像、相互观察、序列、参考文献等。

SCOP 在比较结构相似性时,依据共同结构核心的残基数目及其平均方差的平方根。蛋白质折叠中存在固有的物理和化学规律,这是其分类的基础。构建 SCOP 分类的方法首先是通过自动程序进行视觉观察和结构比较,按照树样层次结构(家族、超家族、折叠、类型)进行整理。这种层状结构中,基本的分类单位是决定蛋白质进化、结构和功能的结构域。作为同一家族的蛋白质,其序列相同性至少为 30%;如果低于 30%,其功能和结构非常

接近,也被列入同一家族(family)。具有相同折叠形式和相似功能的家族归类为超家族(superfamily)。在超家族基础上整理出相同的次级结构折叠(fold),再根据这些折叠分成不同类型(class)。组成 SCOP 的类型包括 4 种主要的球状蛋白质类型,即全 α、全 β、α/β、$\alpha+\beta$,此外还包括其他常见类型,如多结构域蛋白、膜和细胞表面蛋白、小蛋白、线圈绕线圈蛋白、低分辨率蛋白、肽类等。表 6-3 列出了这种层状结构关系和相对数目。

表 6-3 SCOP 层状结构关系*

类型	折叠	超家族	家族
全 α 蛋白	138	224	337
全 β 蛋白	93	171	276
α/β 蛋白	97	167	374
α+β 蛋白	184	263	391
多结构域蛋白	23	28	35
膜和细胞表面蛋白	11	17	28
小蛋白	54	77	116
合计	605	947	1 557

* SCOP1.55 版(2001 年 3 月)收录 PDB 中 13 220 个条目,含 31 474 个结构域

检索 SCOP 时,可按上述树样结构分别在不同层面检索。序列相似性检索方法采用 BLAST,FASTA 或 SSEARCH 的算法。也可按关键词检索,可能得到不同的 SCOP 条目。每个 SCOP 条目链接到结构图像、相互观察、原子坐标、功能变化的构象数据、序列资料,以及 MEDLINE 摘要等。用户可根据 SCOP 检索得到的结果更好地了解感兴趣蛋白质的类别,并有助于了解其进化关系(趋同进化或趋异进化),从而指导功能性实验的设计。

6.8.2.2 CATH

CATH 是 Orengo CA 等[9]于 1997 年建立的另一大结构分类数据库(http://www.biochem.ucl.ac.uk/bsm/cath_new)。其建库基础是根据结构相似性和进化亲缘性,需要人工确认,属于半自动方法。其算法是利用序贯结构比对程序(sequential structure alignment program, SSAP)的结构比较算法并结合动态编程算法(dynamic programming algorithm)。序列比较则是根据配对(pairwise)算法和运用 HMM 方法。与 SCOP 相似,CATH 也包括 4 种树样层次:类型(class)、主体(architecture)、拓扑(topology)、同源(homology)。其中类型位于树型结构的顶端,由次级结构成分组成,目前分为 α 为主、β 为主、α 和 β 三大主要类型以及其他 6 种次要类型。主体结构主要描述次级结构在 3D 空间中的方向性,CATH 中含 30 多种主要的主体结构。这些主体是由约 800 种拓扑结构归类而来的。同源性超家族中,根据结构、序列和(或)功能相似性进行分组,并与进化关系取得一致。根据序列相同性,这些同源结构分成 35%,60%,95%,100% 不同水平,共有约 1 400 种同源超级家族。

CATH 的检索可通过网络服务器进行。用户可得到上述 4 个层面相应的资料,并与

PDB 等进行链接,其应用价值与 SCOP 相当。

此外,DALI(http://www.embl-ebi.ac.uk/dali/domain)数据库也是常用的蛋白质结构比较与分类的数据库。CE 则主要用于三维结构比较。据分析,SCOP、CATH 和 DALI 的分类一致性超过 80%。事实上,不管哪种结构数据库,其可靠性除依赖于所利用的算法外,还依赖于 3D 结构实验数据。全原子接触法(all-atom contacts)是目前常用的一种结构确认的方法。

6.8.3 Pfam

Pfam[10]是一种大型的蛋白质家族的数据库(http://www.sanger.ac.uk/Software/Pfam),截至 2003 年 11 月,该数据库已收录的蛋白家族达 7 255 条目。其资料来源主要为 SWISS-PROT 和 TrEMBL。Pfam 中的每个条目的产生,首先是确定代表该家族的一种种子序列(seed sequence),然后根据该种子序列利用 HMM 模型(hidden Markov model,http://hmmer.wustl.edu)得到多序列比对资料。Pfam 也是 InterPro 协作组的内容之一。Pfam 与许多公共数据库如 SWISS-PROT、InterPro、SCOP 和参考文献等建立相互链接。

Pfam 除含有多序列比对的信息外,也包括结构域、重复区、基序、相互作用等相关信息。此外,还包括一些非结构域的跨膜区、信号肽、线圈绕线圈、低复杂性区域等。Pfam 的检索功能也较强,可通过蛋白名称、结构域、物种分类等查询。

Pfam 在蛋白质空间结构分析、进化遗传分析、相互作用分析、突变研究、药物研究等领域发挥重要作用。

(李 巍)

参 考 文 献

1 Bourne PE, Weissig H. Structure Bioinformatics. New Jersey: Wiley-Liss, 2003. 3~588

2 Krawetz SA, Womble DD. Introduction to Bioinformatics: A Theoretical and Practical Approach. New Jersey: Humana Press, 2003. 565~586

3 Baxevanis AD, Quellette BFF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2nd ed. New York: John Wiley & Sons, 2001. 253~282

4 Lesk AM. Introduction to Bioinformatics. Oxford: Oxford University Press, 2002. 216~262

5 Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers, 1983, 22: 2 577

6 Jansen R, Yu H, Greenbaum D, *et al*. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 2003, 302: 449

7 Callaway J, Cummings M, Deroski B, *et al*. Protein Data Bank contents guide: atomic coordinate entry format description. Brookhaven National Laboratory. 1996, http://www.rcsb.org/pdb/docs/format/pdbguide2.2-frame.html.

8 Murzin AG, Brenner SE, Hubbard T, *et al*. SCOP: a structural classification of proteins da-

tabase for the investigation of sequences and structures. J Mol Biol, 1995,247: 536

9 Orengo CA, Michie AD, Jones S, *et al.* CATH – A hierarchical classification of protein domain structures. Structure, 1997,5: 1 093

10 Bateman A, Birney E, Cerruti L, *et al.* The Pfam protein families database. Nucl Acids Res, 2002,30: 276

第7章 微阵列生物信息学

7.1 微阵列简介

当今生命科学领域中极具发展前景的技术之一是微阵列技术(microarray),常被称为芯片技术(chips),但文献中还是以微阵列的术语为主。微阵列技术得以关注和普及的原因是能同时监测成千上万个基因的表达,并能在同一阵列上对 DNA 或 RNA 进行定量。微阵列充分利用基因组的序列资源来回答这样一个问题:某种特定的物种细胞中,在某一特定的时间、特定的条件下哪些基因得以表达,更重要的是,比较正常和疾病状态下基因表达的不同。这是功能基因组的重要研究内容。DNA 微阵列技术自 1995 年诞生之时[1],就被预计为具有划时代意义的技术,将从根本上改变生物科技的面貌。DNA 微阵列技术的应用数量呈指数级增长这一现象已证实这一预测。例如,在作者写下这段文字的当日(2003 年 9 月 11 日),在 PubMed 中键入"microarray"这一关键词,可得到 4 250 个相关条目。

对微阵列的需求源于基因表达的问题,这是分子生物学的核心问题。尤其是当人类基因组和其他生物基因组序列破译后,常常需要面对的一个问题是:在某种因素影响下,全基因组水平的表达谱是怎样的?例如,对于癌症的研究,尽管许多基因组水平的分析技术有助于阐明肿瘤生物学的某些机制,基因表达谱分析则较以往更有利于理解癌细胞中的遗传改变所导致的一系列结果,即成千上万个基因表达方式的变化。此外,微阵列技术已经成为肿瘤中新的"分子分类学"技术,并用于鉴定诸如人皮肤黑色素瘤等不同肿瘤中新的致病基因。

微阵列广泛应用的另一重要原因是为了理解基因网络(network)或途径(pathway)。传统的分子生物学方法基本上是针对"一个基因一个实验"的设计思路,其通量非常有限。同时,很难获得基因功能的整体框架。例如,传统方法用来解析基因之间相互关系的方法之一是通

过"敲除"某种组织中表达的基因,仅在很小规模上观察对相同或不同组织中对其他基因的表达的影响。这种方法在多个基因功能研究中显得尤为缓慢,效率极低。微阵列可在单一芯片上同时监测整个基因组的变化,因而可同时理解成千上万个基因之间的相互作用,对整个表达谱(expression profile)有一全面认识。这种表达谱可以是来自正常或异常样本。尽管要从所得结果推断基因间的相互作用仍是当代功能基因组中最具挑战性的一项任务,但获得这些资料后无疑可进一步了解该基因所处的途径及其相互作用的信息。另外,一个基因是如何影响另一个基因的表达,有利于阐明某一途径的启动基因,并以此用做药靶。可以肯定的是,微阵列技术将不仅仅在基础性分子生物学的基因研究中起非常重要的作用,也将在疾病机制的了解、诊断和预后,以及药物靶分子鉴定中发挥重要作用。

7.2 微阵列的主要类型

DNA 微阵列是根据传统的 Southern 杂交技术建立的一项新技术。微阵列实验中,标记了的靶分子与已知或未知的 DNA 分子杂交,这些分子被结合和排列到固相支持物上(通常为玻片)。微阵列的分类可根据所排列的 DNA 类型或微阵列的制作方法进行分类。按 DNA 类型可分为寡核苷酸、cDNA 或基因组 DNA。按制作方法,寡核苷酸片段通常是用光刻法原位合成[2]或常规合成后,固相结合到玻璃等支持物上;cDNA 微阵列则是直接用自动控制的针将样品点在固相支持物上[1]。尽管还有其他多种与 DNA 微阵列设计相似的其他微阵列类型,用于检测蛋白质或抗体,目前微阵列领域中仍以下列两种微阵列为主,即美国 Stanford 大学研制的玻片型 cDNA 微阵列和美国 Affymetrix 公司研制的高密度寡核苷酸基因芯片。以下将详细介绍这两种微阵列。

7.2.1 cDNA 微阵列

玻片 cDNA 微阵列是利用自动化装置将 cDNA 或基因组 DNA 文库中的克隆排布于玻片上。与其他支持物如尼龙相比,玻片型 DNA 微阵列有以下几大优点。首先,DNA 样品可以共价结合到已分格好的处理过的玻片上。玻片的透明性和硬性有助于图像采集。其次,玻片可耐受杂交过程中的高温和洗液的高离子强度。再次,玻片的无孔性可以保证杂交体积非常小,以增加靶分子与探针结合的动力学。再其次,玻片的荧光强度低,基本不影响背景噪音。最后,由于待测样品通常是经荧光标记,玻片可保证多种不同的荧光标记分子在同一反应中进行,有利于样品的直接比较。

制造和使用 cDNA 微阵列的过程见图 7-1。其主要步骤如下:①利用 PCR 扩增技术制备 DNA 探针;②利用反转录技术荧光标记靶 DNA;③利用自动化点样装置将 DNA 探针点于玻片上;④将荧光标记的靶 DNA 与探针在玻片上杂交;⑤采集图像数据;⑥数据加工和分析。

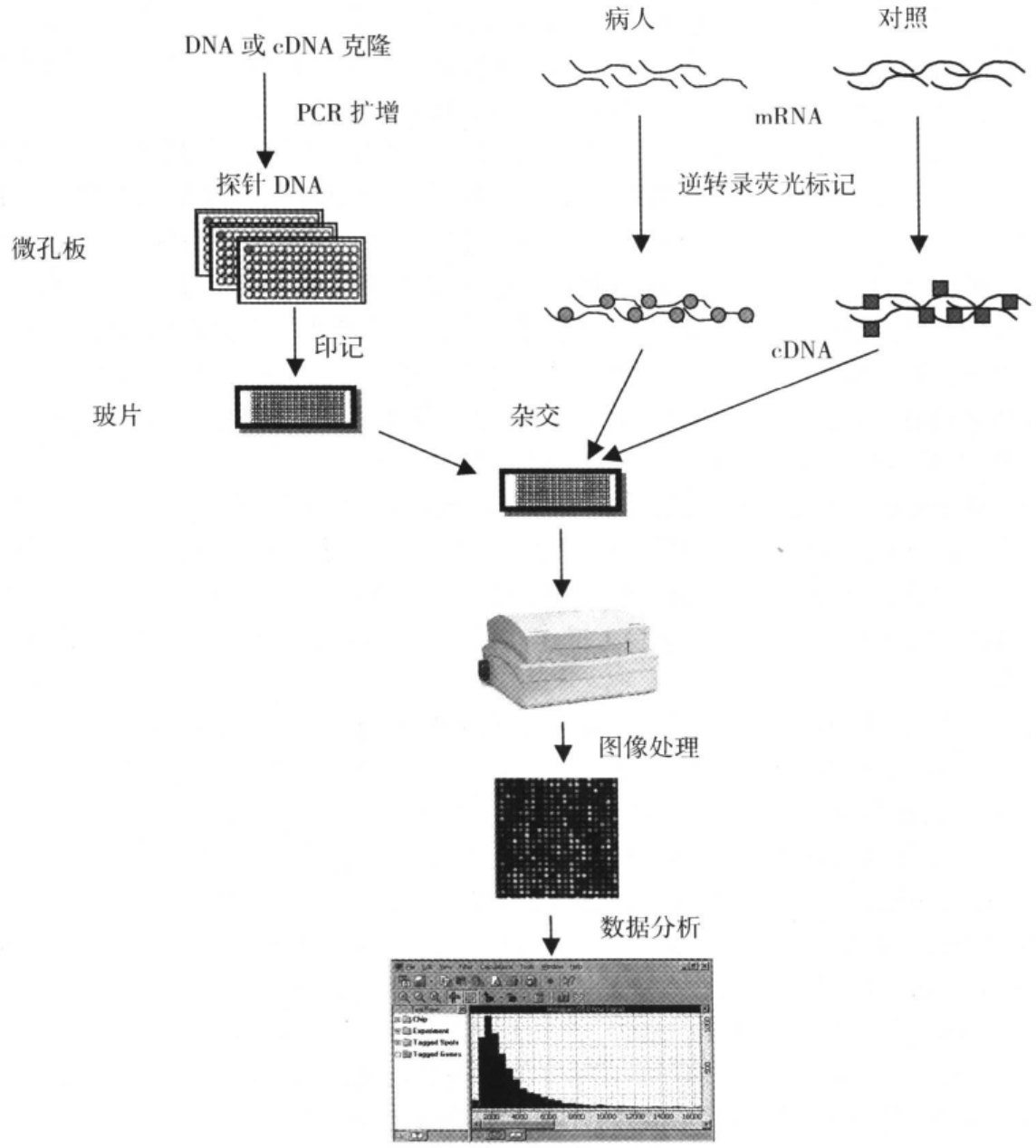


图 7－1　经典 cDNA 微阵列工作流程图

除数据处理与分析分别在 7.5 和 7.6 节叙述外，以下将阐述其详细过程。

DNA 探针制备的常用方法是扩增 DNA 或 cDNA 文库中克隆载体上的 DNA。通常采用一对通用引物进行扩增，便于在相同 PCR 条件下进行。如果是扩增基因组 DNA，则需要选用不同的引物，此时难以统一 PCR 的条件，可以考虑用全基因组扩增技术(WGA)。

微阵列实验中靶 DNA 的制备和标记方法有多种。各种不同方法中，首先是进行待测样本中 RNA 的提取，得到 1 ~ 5 μg 的 PolyA+ 或 50 ~ 200 μg 总 RNA，然后用荧光染料进行标记。最常用的标记方法是通过逆转录过程直接掺入，即在反应中加入一部分 Cy 荧光染料标记的 dCTP，掺入到合成的 cDNA 中。另一种常用方法是间接掺入，即在逆转录反应后，与另一种活化了的染料反应，产生氨基修饰的 dC，该染料较 Cy 标记的 dC 分子小，反应效率

高，可消除前一种标记方法中 Cy3 和 Cy5 不同比例的掺入，以保证靶分子中含相同量的外源基团。第三种也是最少采用的标记方法，是在逆转录反应后，得到单链 cDNA，在荧光标记 dC 存在的条件下，与随机引物反应，进行延伸，得到较短的标记互补链。

微阵列的印制（printing）既是一种艺术，也是一门科学。要得到重复性好的样品点并不容易，需要相关的化学、工程学、编程和分子生物学等方面的经验。尽管已出现多种自制或商品化的机械手用来点样（spotting），其基本制作要求是将玻片放在静止的台面，带有点样针的机械手可沿 x－y－z 3 个不同平面移动排成细密的阵列，还需要有一个清洗站在不同的点样间清洗点样针，同时烘干，其过程受计算机控制。制备过程要避免灰尘的污染，并在合适的温度和湿度条件下进行。制作的最后一步是将阵列表面固定，以使其他 DNA 不能再直接结合到玻片上。已有不同的固定方法，但结果都差不多。通常需要对玻片表面进行一些处理，使其具有亲水性，有助于杂交过程中与溶液的混合。

杂交是指标记了的靶 DNA 与探针在玻片上的反应，产生异源双链 DNA。微阵列上进行的荧光标记 DNA 的杂交条件与其他分子生物学中常用的 DNA 杂交方法相似。杂交反应受多种因素影响，包括温度、缓冲液、湿度、盐浓度、杂交时间、杂交液体积等。杂交反应可以按手工方法进行，也可以在自动化平台上操作。自动化操作在温度控制等方面更好，使得反应的变异性小。

接下来便是微阵列数据处理过程。其目的是通过扫描获得每个 cDNA 点样点的信号强度或信号比值。点样点（spot）与靶 DNA 分子结合后，便带有荧光基团 Cy3 和 Cy5，受适当波长的光激发后，在微阵列上产生荧光信号。产生的荧光信号强度与该点样点的染料分子数目成正比，然后将图像转换为数字信息，记载着基因表达的定量信息，用于后续的微阵列数据分析。微阵列图像处理的过程包括：①微阵列目标区的分割；② 背景荧光强度的获得；③ 目标区检测；④目标荧光强度的获得。目标区的分割是通过软件来设定要检测的微阵列点所在的区域，以获得该区域内各样品点的荧光信号强度。背景分析是分别得到检测点附近的背景强度用于校正。玻片上的背景强度并不很均一，因此不能只采集固定位点的背景强度。背景的变化往往是渐进的和平滑的，可能受多种技术因素影响。突发性的变化很少见，一旦发生，其附近区域的样品信号强度往往不可靠。目标区测定要求在指定的区域内进行。最后一步是计算各点样点的荧光强度的数字信息、背景强度和质控信息。

7.2.2 高密度寡核苷酸阵列

这类微阵列与上述的 cDNA 点阵完全不同。高密度寡核苷酸阵列不需要 cDNA 克隆，因此制备时不需要点样过程，而是将寡核苷酸（通常为 20～25 mers）在平坦的固相支持物上固相合成。各个寡核苷酸链是通过前一个碱基的 5’羟基与下一个待连接碱基的磷酸基团进行共价结合。每个待结合到寡核苷酸链的核苷酸的 5’位置受到保护，以免在每一轮合成反应中一个以上的碱基被掺入。保护基团在下一轮反应前，在酸性或光照条件下转换为羟基。这种去保护方法的不同导致不同的原位合成方法的出现。最著名的是 Affymetrix 的 GeneChip（基因芯片）技术和 Agilent 的 Ink Jet（喷墨）技术[3]。

GeneChip 利用光刻面罩和光去保护法进行寡核苷酸的合成。首先将一种带有可利用光化学切除的保护基团的连接器（linker）贴附在玻片上，然后利用光刻面罩（mask）通过发

光激活阵列区,产生活性羟基。将特定的碱基(A、C、T 或 G)在给定条件下与活性表面温育,产生共价连接反应。其制作过程通常需要 4 个掩饰步骤(参见图 7-2)。然后按待合成的寡核苷酸序列依次掺入,重复上述步骤,得到不同的寡核苷酸探针。一套完整的探针阵列,当长度为 n 时,需要 $4n$ 个合成步骤。该技术可制作密度非常高的阵列。尽管合成步骤中需要不同的面罩,当设置后,其合成就变得简单,可生产出大量相同的阵列,便于标准化和商业化生产。

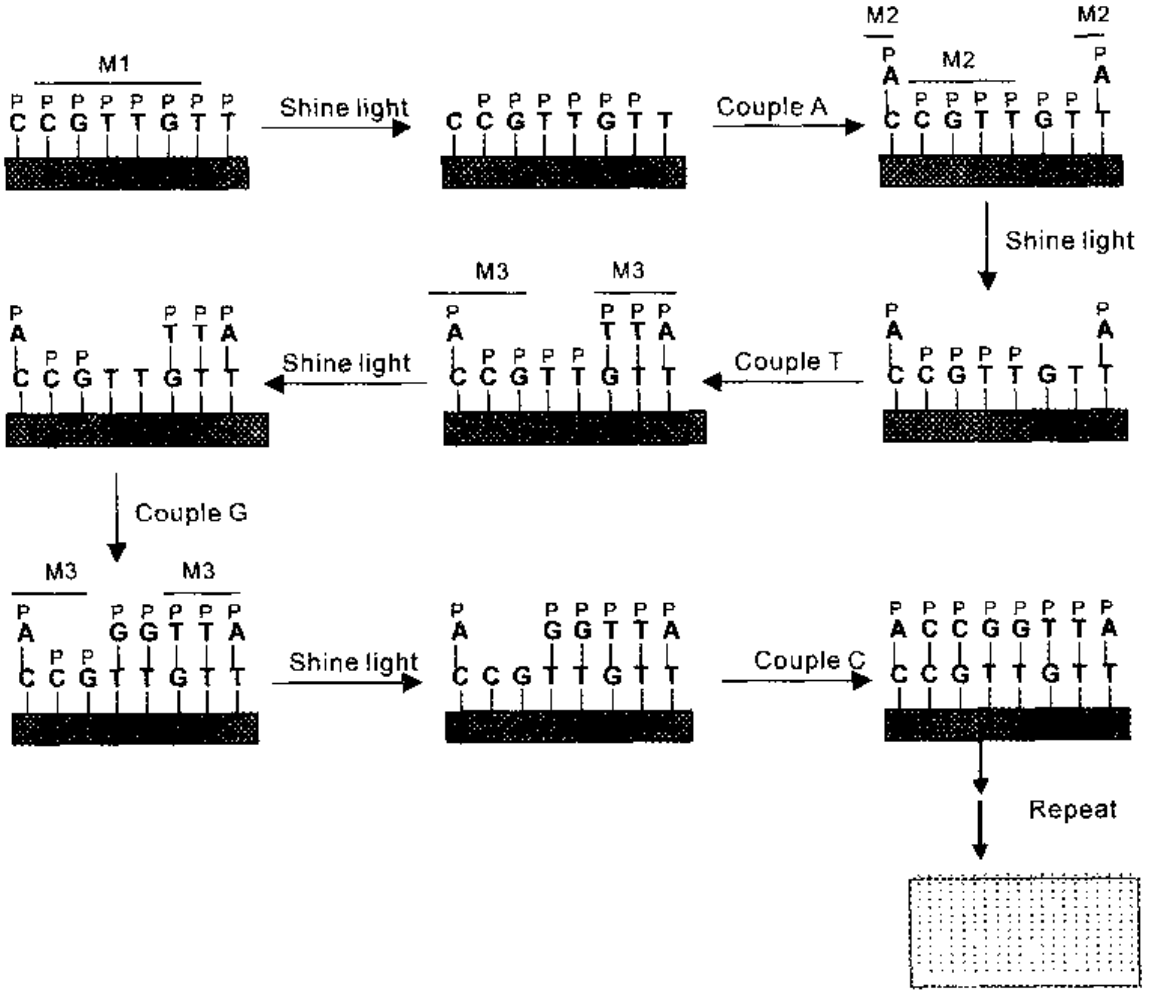


图 7-2 微阵列光电法制作示意图

Ink Jet 技术利用了与喷墨彩色打印机相同的原理。该技术不用光进行保护基团的转换,去保护过程利用 DNA 合成仪中相同的化学方法。它需要 4 种不同的"墨盒",分别装载着 A、T、G、C 4 种不同的核苷酸。合成的每个步骤中,程序指定的碱基通过喷嘴被喷到玻片上指定的阵列点,打印头按阵列所需的序列要求"打印"出特定的核苷酸,进行合成反应。整个输出过程由程序控制,较 GeneChip 更为灵活。图像和数据采集都采用相应的商用软件。

寡核苷酸阵列可得到比 cDNA 阵列更高的密度,因而可利用冗余探针作为对照来检验结果。探针通常分两种,即完全匹配(perfect match, PM)和不匹配(mismatch, MM)。但 cDNA 微阵列的优点在于其灵活性,可根据用户需要进行设计,而且样品类型多样。表 7-1 比

较了两者的优缺点。

表 7－1　两种微阵列的比较

cDNA 微阵列	寡核苷酸微阵列
制作相对便宜	价格昂贵
探针序列长，特异性高	探针序列短，特异性较差
灵活性好，可采用各种序列	仅限于给定序列
点样均一性较差	样品更均一
通常自行制作	可直接购买
系统变异性大	数据更可靠
数据分析更容易	数据分析复杂

7.3　微阵列的应用

微阵列已成功用于不同领域，这取决于微阵列实验的类型。微阵列最常用于转录组的研究。早期大部分的表达谱实验是用来比较两种样品中表达的差异性[4,5]。在设定的界值为(cut off)n 时(n 为倍数)，那些比值高于 n 的基因被称为差异表达基因。目前则主要利用统计学方法处理后判断是否出现表达差异，而不是依靠某种人为设定的界值来判别。最常用的两种统计方法是 t 检验和方差分析(ANOVA)，前者主要分析两组数据的变化，后者主要用于多组数据的比较[6]，参见第 10 章。

表达谱检测的另一重要用途是通过观察不同时相多个样品表达谱的变化，以预测基因的功能及其所参与的生物过程，或对样品的表型和组织类型进行分类。这种预测是基于这样一种假设：具有相同变化方式的基因隶属于同一途径；具有相同变化方式时其样品类型非常接近。如果已知某些基因和(或)样品的功能或类别，具有相似表达方式的其他基因很可能参与同一细胞过程，或具有相似的样品特征。这是相关性的证据。将转录物和样品分类是依据共调节作用，或依据表达谱将细胞所处的状态进行分类。

微阵列技术具有应用前景的领域之一是肿瘤的分类。微阵列可同时监测原发性肿瘤或细胞株中成千上万个基因 RNA 表达水平的变化。癌症的起因是由于遗传或外遗传(epigenetics)变化的累积效应，这主要由于癌症相关基因的变化所引起，如癌基因或抑癌基因，或参与细胞周期控制、凋亡、黏附、DNA 修复、血管形成的基因等。组织学上难以区分的肿瘤，其临床特征可明显不同，这些肿瘤在分子表达水平的亚分类将有助于解释肿瘤对治疗反应的不同。例如利用微阵列技术，根据“邻近分析法”和所得到的肿瘤预测模型，用于对白血病的分类，可较好地区分急性髓性白血病和急性淋巴细胞性白血病[7]。类似的方法应用于 B 细胞淋巴瘤[8]和皮肤黑色素瘤[9]的分类。这些例子说明利用微阵列技术可鉴定出以前分

类不明确但临床意义较大的亚类。

类似用于癌症的分类,微阵列技术也可用于药物的研发。最简单的应用是观察用药后肿瘤细胞株或肿瘤病人基因表达谱的变化。这些研究可提供药物作用的精确机制或鉴定出药物反应的早期生物标记。对于长期药物治疗的临床观察,微阵列可得到有关副作用的信息和用于预测副作用的生物标记。微阵列技术的其他用途包括:①鉴定疾病的易感基因;②建立有效的和早期诊断的指标;③ 通过基因表达谱的监测建立可靠的预后指标;④根据患者基因表达谱作为个体化治疗的依据。此外,微阵列基因表达谱还可用于测试药物或一些化学物对不同生物体的作用,用于测序、SNP 检测或遗传印记(亲子鉴定)分析等。

微阵列技术应用虽然非常广泛,但也有其局限性。首先,该技术对于低表达基因的检测灵敏度有限。其次,基因表达谱不能反映翻译后修饰(如磷酸化)和蛋白水平的真实变化,如翻译启动和降解的变化。还有一个缺点是由于不同探针间的相似性得到相互杂交信号,出现假阳性干扰结果,这在cDNA微阵列中特别明显,各点样点很难避免与其家族成员的交叉反应。最后,微阵列实验的结果不能用于证实某些生物变化,而只是通过统计分析得到一个可能的推断,需要进一步用其他传统的分子生物学方法等确认。

7.4 微阵列实验设计

自从上世纪 90 年代中期微阵列技术得以应用以来,已有许多综述文章讨论该技术及微阵列的数据处理,尤其后者是该技术的一大挑战。但是,有关微阵列实验的设计很容易被研究人员所忽视。许多生物学者认为,实验中只要包括实验组、对照组和重复实验,已足以回答所要解决的问题。在获取相关的实验资料后,才考虑数据分析问题。通常出现的问题是实验所提供的所需数据不足以解决相关问题,特别是具有数十个或数百个微阵列的实验,尤其如此。多数微阵列实验是用来比较两组或两组以上样品中差异表达的基因或寻找多个样品中某些基因表达方式(pattern)。多个来源产生的噪音是获得准确资料所主要关心的问题,因而影响到所要解决的问题。消除这些噪音是微阵列实验设计的主要动因。

恰当的实验设计,可使噪音降到最低。这需要研究者审慎考虑各种实验影响因素,并鉴定导致噪音的原因。一种常识问题是不同的研究项目需要根据所需解决的问题提出不同的实验设计,但微阵列实验设计中,不管分组数目和样本数目的多少,仍有以下一些共同的问题需要考虑。

7.4.1 对照组的选择

微阵列实验中,使用对照(control)主要有两种目的。一种对照是为了数据的正态化,如cDNA 微阵列中使用看家基因和封闭 DNA(spike DNA)(详见下文)。另一种对照具有更广泛的意义,就是设立阳性和阴性对照,以便对结果有信心,并监控实验是成功抑或失败。cDNA 微阵列常用的一种对照是与实验组的对照标本,即用一种荧光染料标记一种 mRNA 样品(如正常组织),而用另一种染料标记另一种组织(如癌症标本)来源的 mRNA。这时需观察的组织如癌组织,又称处理组(treatment),正常组织被称为对照组。同样当做处理组的还

可以是暴露于药物、病原体或环境因素(如热休克或辐射的标本)中的组别。在所有这些处理因素中,尽量选择除处理因素不同外其他因素都相同的参照系统作为对照组。

在如何使用参照样品的问题上仍存在争议。一些研究者认为如果将所有标本都与同一参照系比较,可导致实验可靠性的不足。多数微阵列实验中的参照标本使用与实验组相同性能的染料。实验设计中,最重要的考虑是对照样品量足够、具有均一性、且稳定性好。多重实验中,对照样品不一定与实验本身有关,只是尽可能使基因测定时产生非零信号。这是因为当对照用做分母时,如果为零或数值非常小,所得到的实验/对照的比值难以计算或波动很大。参照系的选择还常采用零时点进行比较。通常癌症标本表达基因的数量较正常组织多。各种不同的参照系,具有各自的优缺点。零时点mRNA对照设置相对方便,并容易阐释比值,但标本量大时不容易做到,而且不便于不同实验室之间的比较。肿瘤细胞株 mRNA 常被用做不同实验室间的共同对照。采用基因组 DNA 对照也很便利,量足,而且重复性较好,但对于高表达的基因,这种对照所得到的分母相对较小。其他对照标本还采用与实验组别相同,但将不同组织的 RNA 进行混合(pooling),以保证实验组产生的基因表达信号,对照组中都有等同基因的表达。

7.4.2 重复样本的使用

早期 cDNA 微阵列实验通常是两种不同荧光标记的 mRNA 样本与同一芯片上的探针竞争性杂交反应。这种比较性的实验方案似乎无须进行重复。其实不然,也会出现许多问题。如前所述,误差变异可来源于染料相对强度的偏差,或者染料对 DNA 的亲和力,以及 RNA 提取和处理过程等。光凭一个微阵列实验的结果很难评估这些变异性,因此需要重复实验来对这些误差进行适当控制,以反应基因表达的真实情况。

重复(replicates)实验的设计可以是将同一个基因点在不同的点样点上,或者用不同的微阵列重复某套基因。前者主要解决点样部位不同所引起的噪音,后者主要解决杂交过程或生物学变异引起的误差。

重复实验的重要性可用以下例子加以说明。假设两块微阵列片上不表达基因(即背景噪音)的荧光强度符合正态分布,均数(μ)都为 1 000,标准差(σ)都为 200。在一块片子上,每个基因只点 1 个点,另一块片子上每个基因各点 4 个点。一种不表达基因在测定值高于一定量(例如 1 400)时,被当做表达基因的概率的计算方法如下。

对于单个点的微阵列片的 X 变量值(这里 $\bar{X}=1\ 400$)先按下式转换为正态变量(Z):

$$Z=\frac{\bar{X}-\mu}{\sigma}=\frac{1\ 400-1\ 000}{200}=2 \qquad \text{(公式 7-1)}$$

假阳性的概率为(查标准正态分布表):

$$P(Z>2)=1-P(Z\leqslant 2)=1-0.977\ 2=0.022\ 8 \qquad \text{(公式 7-2)}$$

结果表明,有 2.28% 的机会不表达基因被误做表达基因。乍一看来,这个数值似乎很

低，但是，一块 cDNA 阵列上通常有多达 2 万个点。如果其中 50% 的基因不表达，其假阳性基因数目可达 200 个。这个数目甚至多于那些实验中真正有表达差异的基因而混淆实验结果。

对于另一块每个基因同时点 4 个点的片子，即每个基因被测定 4 次，其标准差的计算为：

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{4}} = 100 \qquad \text{（公式 7-3）}$$

根据此标准差，再转换为正态变量为：

$$Z = \frac{\bar{X} - \mu}{\delta_{\bar{x}}} = \frac{1\,400 - 1\,000}{100} = 4 \qquad \text{（公式 7-4）}$$

这种情况下，假阳性的概率为：

$$P(Z > 4) = 1 - P(Z \leqslant 4) = 1 - 0.999\,968 = 0.000\,032 \qquad \text{（公式 7-5）}$$

以上结果表明，不表达基因只有 0.003 2% 的机会被当做表达的基因。同样，若具有 2 万个点的微阵列所代表的基因数目为 5 000 个，如果 50% 的基因是表达的，最多只有一个基因（0.16 个）是假阳性。如果用 4 块这类片子检测总计 2 万个基因，所得到的假阳性率总和为 0.64（4 × 0.16）。这表明如果重复实验结果显示某个基因具有高表达值，则该基因极有可能是表达基因。

毫无疑问，不论是在同一芯片或不同芯片上进行重复点样时，对于减少假阳性和假阴性非常重要。从统计学观点来看，重复越多越好，但由于成本的大大增加，往往不是很现实，而且会增加许多没有必要的数据处理的负担。选择重复数目以得到统计学上有意义的结论取决于多种因素，包括预期基因变异性大小、出现Ⅰ类错误概率（通常为 0.05 或 0.01）和统计效力的界值、对照和处理组均数差异性大小等。重复数目与标准差成正比，而与某种固定的Ⅰ类和Ⅱ类错误概率的条件下的两组均数的差值成反比。

尽管选择基因重复数目受多种因素影响，一些研究者常凭经验选择合适的数目，无论如何有重复比没有重复要好。一般建议每个基因最少需要 3 个点，位于 3 块不同的微阵列片上[10]，这是由于 3 种类型可得到三角关系的结果。

7.4.3 随机化原则

微阵列实验设计中另一种常用方法是随机化原则（randomization），以减少实验偏差。随机化可在不同水平进行。最低层面是重复点的随机化。如果重复点彼此靠近，该区域的局部偏移将无法判别该结果是由于基因的效应还是芯片本身的效应。这就要求微阵列制作时考虑随机化原则。这也给数据采集和资料分析带来一定的困难，特别是那些非标准化的芯片。解决方法是固定某种模式的区间内进行重复样品的点样。较高水平的随机化是将芯片随机指定用 Cy5 还是 Cy3 染料。更高水平的随机化是在处理组和对照组分组时进行随机

抽样。

7.5 数据处理

7.5.1 丢失数据和极端值的处理

丢失数据(missing data)和极端值(outlier)是微阵列实验中数据质量控制(quality control)的两个基本问题。数据丢失的原因很多,包括分辨率不够、图像失败或只是由于芯片上的灰尘或划痕所引起。数据丢失还可由于自动化方法中的系统误差。多数情况下,丢失的数据是这些不同原因相混合,不存在哪种占主要的问题。cDNA 微阵列中数据丢失的含义是由于空点(empty spot),其荧光强度为零,或者由于其背景强度高于样品点。Affymetrix 微阵列丢失数据是指原始数据中错配值(MM)高于全配值(PM)。这些可疑资料通常是经过手工方法剔除,不用做分析。但在某些情况下,剔除丢失数据可能给数据分析带来困难,并引起重要信息的缺失。因此,最好将丢失数据进行替换。最简单的数据替换方法是根据同一芯片上其他点的情况进行统计分析而得到一个预计值。对于双色 cDNA 微阵列,如果某个基因有重复点,这些点的平均值可用来代替丢失数据。如果没有重复点,可用统计方法预测丢失数据(如 EM 算法)。一种简易方法是计算该样品点用不同染料标记时在整个芯片强度的分布位置,并以此为参照,推算出相应位置上的丢失值而加以替换。

极端数据是指那些偏离群体的数据。微阵列实验中,极端值的出现和消除可在不同水平。极端值可在一块芯片上出现,但重复片子上不出现;也可以是同一片子上某个基因的重复点,而不管这些重复点邻近与否;还可以是同一片子上任意点所产生的偏离。

现有微阵列技术中,多种因素可导致不同芯片间的变异性。已有不同方法减少这些芯片间的变异和系统误差(如下文将要叙述的正态化)。同一类型的芯片中,那些变异性大的片子应当去除,这种片子又称极端片子(outlier slide)。片间变异可能由于点样浓度和体积、加到芯片上的标记靶分子数目、杂交条件和其他因素等所引起。最简单的去除极端片子的方法是靠视觉观察图像。一种简单而有效的消除方法是通过提高实验自动化程度而消除。

另一种去除极端片子的方法是如前面实验设计中讨论的那样,进行重复性实验,并用统计方法评估片间变异。重复片子上对应的基因可得到相关系数。这种方法中,至少需要 3 次重复才能评估芯片质量和剔除极端片子。所用的方法是计算两两配对(pairwise)相关系数。这时,需要设置一个相关系数界值,依实验设计而定,但通常必须大于 0.9。通过两两配对,分别得到各相关系数值。通常情况下,相关系数都比较高且差别不大。如果两个相关系数值远远低于另外一个,常表明存在极端片子。如果所有相关系数都很低,表示微阵列的质量差,这不是极端值的范畴,而需要重新设计和制作芯片。

同一芯片上也可出现极端值。在缺乏重复点的情况下,最高和最低的数值通常被当做极端值处理而加以剔除。阈值的设置可以根据百分位值(如最低数值点或最高数值点的 0.5%)或那些偏离整个片子的分布中一定数量的标准差范围(如 $\pm 3\sigma$)。剩余的资料重新计算均数和标准差。这个过程不断重复,直到没有发现极端值为止。这种方法主要根据统

计学原理,有一定的局限性。从生物学角度来看,某些基因的表达可能极高或极低,而且意义很大,特别是那些高表达的基因。在有重复点的情况下,极端值的挑选主要根据重复性的相似情况。如果某个重复点偏离该基因所有重复点平均数几个标准差范围,这个点即被当做极端值,需要去除或替换。

7.5.2 数据的正态性和线性检查

正态性(normality)是指所分析数据是否符合正态分布,而线性(linearity)是指两组样品的散点图中,其数据相关性呈线性。在数据分析前后都必须检查数据的正态性和线性,这是由于微阵列数据分析所用统计方法中基本都假定数据呈正态分布。如果数据不呈正态分布,而是向一侧偏移,这些统计方法所得结果将不可靠,除非选用不依赖正态分布的非参数统计方法。微阵列数据通常向右歪斜,这是因为许多基因表现为中度或低度表达。检查正态性的常用方法是进行 Q-Q 作图(图 7-3a),比较测试样品累积分布函数和正态性密度的累积分布的关系。如果样品分布接近正态,其图形在线内分布,依据偏差方式可以知道其偏移方向。如果曲线上各点表现为底部较顶部平坦,这时称数据左歪;相反,如果曲线上的点表现为顶部较底部平坦,则称数据右歪。

数据的线性检查有助于正态化方法的选用。线性相关资料可用整体正态化方法,非线性相关资料则采用局部正态化方法(详见下节)。同时,线性检查也可提供数据可靠性的信息。检查线性最简单的方法是作散点图(scatter plot)。如果数据呈线性,点的分布应符合直线性(图 7-3b)。

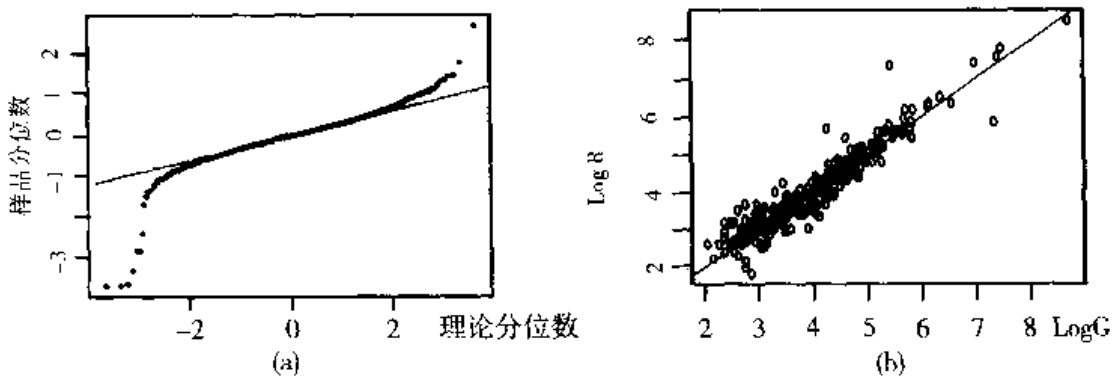


图 7-3 Q-Q 作图

(a)Q-Q 作图比较样品分位数与理论分位数,偏离线的数据点表示不符合正态分布。

(b)对数转换值散点作图,显示数据的线性关系。

为了提高数据的正态性,通常将微阵列原始数据在数据分析前进行对数转换。数据转换后可更好满足统计分析性能,如变异性的稳定和获得累加性。对数转换不但可将数据转换为两个通道中所得数据的比值,同时也使其变异性稳定下来,并将数量级差别转换为累加性差别,有利于统计分析。

7.5.3 数据标准化

7.5.3.1 什么是标准化

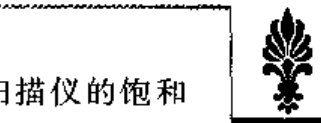
微阵列技术使得研究者可同时监测成千上万个基因的表达情况。但是微阵列实验中会出现许多不合需要的系统误差[11]。表7-2总结了cDNA微阵列实验中常见的误差来源。

表7-2 微阵列实验误差来源

误差来源	评　　价
mRNA样品制备	因组织类型、试剂盒、对RNA降解的敏感度不同,样品间的靶分子可出现差别
标记方法	受标记方法、操作过程及核苷酸组成的影响
扩增方法	PCR扩增难以避免扩增量的差别
点样针	点样针的表面性能有差别
探针体积	即使用同一点样针也会出现移液量的细微差别
探针固定	尚不清楚cDNA探针中究竟有多少量被结合到玻片表面
杂交反应	受许多参数的影响,如温度、时间、缓冲液和靶分子数量的影响
玻片的不均一性	可能因为表面封闭的不均匀,导致探针分布或杂交反应不均一
非特异性杂交	无法完全消除的一种常见误差
非特异性背景	非特异信号的产生可来自邻近点的信号或背景的平均强度的加深
图像分析	非线性传递特性、饱和效应和点形可导致误差
图像分割	光是一种污染信号
点形	点形不规则很难从背景中分割出来

标准化或正态化(normalization)就是将不同芯片中的差异进行消除的过程。例如,分别代表对照组和处理组的微阵列中,两者的差异可能来自起始RNA或cDNA量的不同,或者标记反应的效率不同,抑或杂交过程的不同。其中任何一种因素都会使结果偏移而影响基因表达水平的测定。因此,标准化在微阵列数据分析的前期阶段非常重要,主要减少非生物性影响因素的影响。

标准化是进行数据准确分析的关键。目前已有许多针对cDNA和高密度寡核苷酸微阵列标准化方法。同时也建议其过程要尽量避免真实生物学数据的丢失。并不是标准化方法越复杂,数据就越可靠。标准化方法的选择取决于实验设计、数据特征等,更重要的是考虑引起误差的主要来源。因此,标准化的首要问题是认识系统误差的来源。如果采用参照样本,微阵列中的误差来源容易寻找。例如鉴定荧光强度依赖性误差的常用方法是将比值的对数值与荧光强度对数值的平均数作散点图(即M-A作图)。对于双通道cDNA微阵列,处理组和对照组的散点图应呈直线。M-A作图中,多数点应沿水平线分布,即由于多数基因的表达不表现出差异。数据从水平线的大量偏移常提示存在某种误差。图7-4显示荧



光强度依赖性误差与相应的潜在误差的来源。其中高强度端的切断表示扫描仪的饱和(A),低强度端的高变异性表示通道特异性叠加性误差(B),高强度端的高变异性表示通道特异性乘积性误差(C),曲线(D)表示通道平均背景变异性,曲线(E)表示斜率的变异,MA图形分离表示不均一性(F)。(http://www.jax.org/staff/churchill/labsite/pubs/index.html)

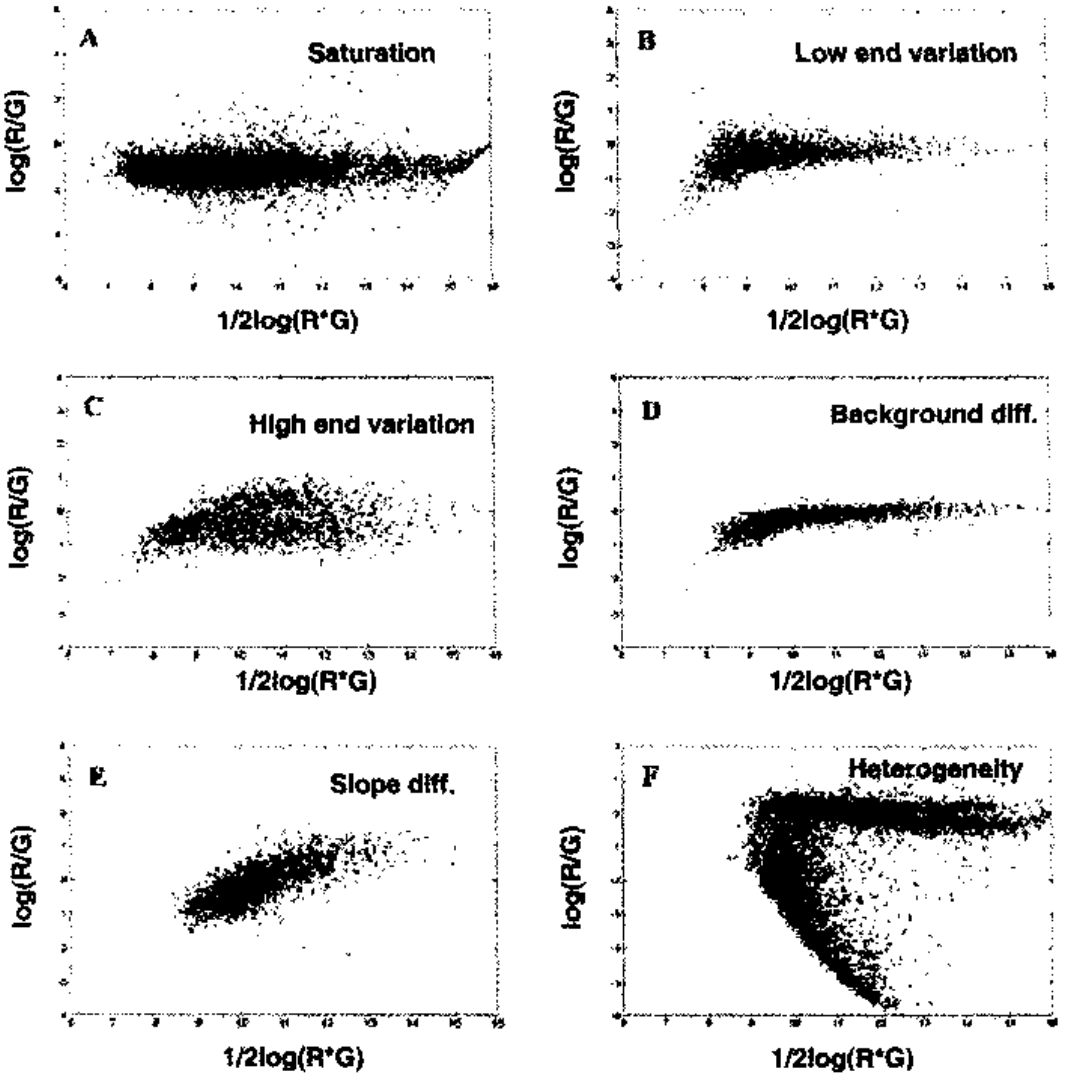


图7-4 强度依赖性变异来源

7.5.3.2 如何进行标准化

从统计学角度,微阵列数据的标准化是通过不同的数据转换方法以便不同样品或实验具有可比性。最基本的正态化方法是将荧光强度或强度比值进行对数转换以达到接近正态分布,从而减少荧光强度或其比值的变异性,使其不受绝对值中数量级差别的影响。另一种数据转换过程,也就是通常意义的标准化过程,是将来自不同实验的数据转换后具有可比性。最常用的方法是将数值转换为 Z 值。微阵列数据处理中最常用的数据转换方法又称中心化,就是将某个分布移到预计均数处使之对称,便于两个或多个样本的比较。所有这些方法的最终目的是将数据正态化,使得各基因相应的数值间具有可比性。实现这种转换的方法大致分为两大类:整体正态化(global normalization)和局部正态化(local normalization)。

整体正态化包括总的荧光强度正态化和对照基因的使用，通常在两样品呈线性关系时采用[图7-5（a_1）]。总强度正态化基于这种假设：当该生物的基因组都印在微阵列上时，两个这种样品的基因表达大多数相同。特别是当样本量足够大、来源也相似时，更是如此。如果样本来源不同，这种正态化效力降低。整体正态化中涉及不同的方法。其中之一是将每个芯片的平均强度调到同一水平[图7-5（a_2）]。这种方法的优点是将中心平均后，数据不会扩散，即原始数据的信息量没有改变。这可以通过将芯片上所有值都除以该平均数而达到目的。如果分布出现噪音或偏移，可用百分位数、中位数等代替，这些数值受无关信号影响小。另外也可以用平均减法来进行中心平均，与除法原理相似，常用于对数转换资料，见下式。

$$\log X_{gene} - \log Y_{mean} = \log (X_{gene}/Y_{mean}) \qquad (公式7-6)$$

另一种方法的假设是，微阵列上所有的基因其变异性相同。这样，整体正态化就是建立一个线性回归模型，使之符合直线形式：$y_i = \alpha + \beta x_i + \varepsilon_i$。这样处理的资料，其误差符合最小方差的原理。这是对最佳直线的拟合，使得两组数据具有相同水平的总强度。

其他整体正态化方法还采用封闭 DNA 对照或看家基因对照进行的校正。封闭 DNA 对照中，假定其强度比值为1，如果比值偏离1，可用下式来调整样品间的差别：

$$调整强度_{cy3} = 测定强度_{cy3} \times (封闭对照强度_{cy5}/封闭对照强度_{cy3}) \qquad (公式7-7)$$

看家基因方法与此相似，假定细胞之间看家基因的浓度是相对固定不变的，事实上并不一定如此，这种校正可能是误差的一个来源。

局部正态化用于两个或多个样本之间呈非线性关系的情形[图7-5（b_1）]，这通常是由于系统误差所产生，包括点样针的不一致、玻片表面的不均一、局部杂交条件的不一致等。如果能解析出微阵列的亚类（subarray），可以利用上述整体正态化方法处理这种亚类，从而消除这种系统误差。另一种局部正态化方法是被称为 LOWESS 的局部权重多项回归法。这种方法将数据拆分为彼此重叠的区间，然后利用下列多项式进行拟合：

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots \qquad (公式7-8)$$

LOWESS 回归通常在 A 对 M 的区间内进行，这样低或高表达的基因很容易从散点图和线性关系中发现（图7-4）。LOWESS 正态化的处理过程是：先将数据区分为给定大小的不同区间，落在该区间内的点属于多项式的首项并进行加权，也就是说更靠近估计点的数据权重要大。这样所得到的结果是一条平滑曲线，它与局部多项式点数成正比[图7-5（b_2）]。

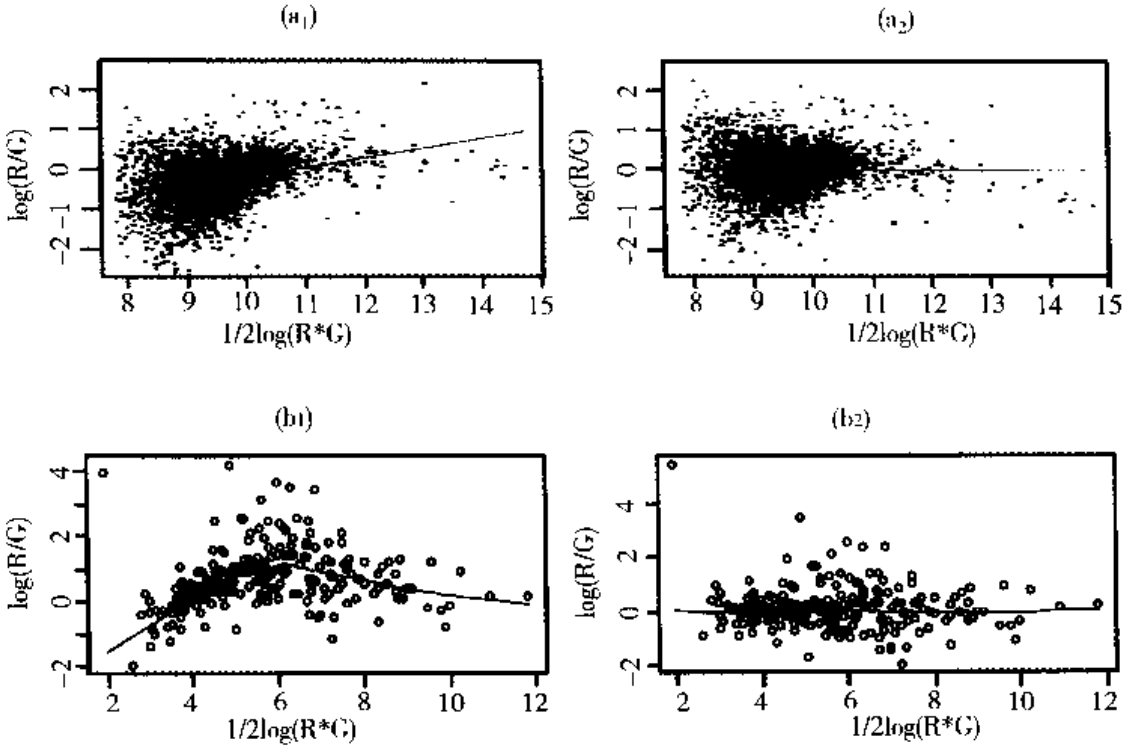


图7-5 双通道样品正态化前后的M-A作图

(a_1)正态化前的线性相关资料;(a_2)总体正态化后结果;

(b_1)正态化前的非线性相关资料;(b_2)利用LOWESS方法局部正态化处理结果

7.6 数据分析

微阵列合适的数据分析方法取决于实验设计和研究目的。微阵列通常用于发现基因、阐明代谢途径和进行分子分类。要回答这些问题,需要综合考虑多个方面。实验条件可以有多种,如不同时相观察。这些条件可以相对独立,也可有某种联系,甚至是多种实验变量的组合。为了反映这些多样性,已有一些常用的方法用做鉴定这些变化的显著性或基因表达模式的识别。这主要包括监测两个或多个样品基因表达水平比值的表达差异性(differential expression),减少维数并进行归类的主成分分析(principle component analysis),以及用做类型发现(class discovery)和类型预测(class prediction)的聚类分析(clustering)和分类分析(classification)。以下将主要介绍目前常用的差异表达分析和聚类分析。

7.6.1 差异表达分析

差异表达的基因是指那些低或高表达的基因。例如,与正常组织相比,肿瘤组织中相对高表达的基因。微阵列技术的早期应用中(目前还在应用),研究差异表达基因的方法是将相同组织来源的两种样品(如癌症和正常)经不同标记,混合后与同一芯片杂交。筛选的标准通常定义为1.8~2.0倍。其比值超过这个界值时被认为是差异表达。如果使用重复点,

出现一个以上的重复点的表达比值超过阈值，这个基因可被认为是差异表达基因。

这种倍数变化分析方法具有明显的局限性，这是因为缺乏基因表达变化数据的可靠性和界值的选择标准。从生物学角度看，基因表达变化的程度并不一定表示会产生生物学后果。而两种不同组织或状态下，一种仅表现为20%变化量的基因可能较同样组织中变化量超过2倍或更高倍数的基因更具生物学意义。例如，信号传导途径中，任何细小的基因表达量的变化可产生明显的生物学结果。即使通过统计分析得到的差异表达基因，也会出现同样的争议。另外，低表达基因的荧光强度更易受到其他因素如背景噪音的影响，因此低丰度基因所受影响较高丰度基因大，需要一个更大的界值才能筛选出受调节的基因。

筛选差异表达基因更可靠的方法是利用统计学原理，特别是有重复芯片或基因点时。仅有为数不多的统计处理方法可用来进行微阵列的数据分析。差异表达基因的统计检验方法通常是比较两组或多组均数的差异。如果仅有两组，可用 t 检验；两组以上则常用方差分析(ANOVA)，两者的假设都必须符合正态分布。其统计检验方法参见第10章。

差异表达统计分析时，需要决定选用单侧检验还是双侧检验。通常分3种情形：①研究组(如肿瘤)较对照组表达高；②研究组较对照组表达低；③研究组和对照组的表达可高可低。前两种情况选择单侧检验，最后一种情况选择双侧检验。另一需要考虑的问题是界值 α(Ⅰ类错误)的设定，通常选择0.05。由于微阵列上有成千上万个基因，尽管Ⅰ类错误的比例较小，但假阳性的基因数目不可低估。如10 000个基因的芯片，将有500个基因的表达为假阳性结果。这种错误率显然与样品大小有关。基于这种分析，很难避免实验水平上的误差，这可以根据下面的计算进一步说明问题。如果选择显著性水平为0.05，每个基因不出错的概率为：

$$P_{gene}(\text{正确}) = 1 - p = 1 - 0.05 = 0.95 \qquad (\text{公式 } 7-9)$$

因此，如果观察500个基因，实验水平上不出错的概率为：

$$P_{exp}(\text{正确}) = (1 - p)^{500} = (1 - 0.05)^{500} = 0.95^{500} = 7.275\text{E} - 12 \qquad (\text{公式 } 7-10)$$

这样，实验水平引起错误的概率是：

$$P_{exp}(\text{错误}) = 1 - (1-p)^{500} = 1 - (1 - 0.05)^{500} = 1 - 0.95^{500} = 1 - 7.275\text{E} - 12 \approx 1 \qquad (\text{公式 } 7-11)$$

就是说，含500个基因的微阵列将不可避免地出现错误，更不用说含有上千或上万个基因的情况。因此，多重比较时常需要进行校正，以降低总的实验水平的Ⅰ类错误的概率，即在基因水平发生至少一个错误的概率。校正方法中常用的有 Šidák 校正法[12]、Bonferroni 校正法，以及假发现率控制法(false discovery rate controlling)[13]和置换校正法(permutation correction)。

多重比较进行校正的目的是为了控制在实验水平犯Ⅰ类错误的几率，这个几率等于在

基因水平犯至少一个同样的错误：

$$P_{exp}(\text{错误}) = 1 - (1 - p)^R \text{ 或 } \alpha_e = 1 - (1 - \alpha_g)^R \quad \text{（公式 7-12）}$$

其中 α_e 表示实验水平上预期 I 类错误概率（如 0.1），α_g 表示基因水平 I 类错误概率，R 表示基因的数目。

Šidák 校正法的表达式为：$\alpha_g = 1 - (1 - \alpha_e)^{1/R}$。Bonferroni 校正法的表达式为：$\alpha_g = \alpha_e/R$。这两种校正法都有其缺陷性。因为当基因数 R 很大时，所要求的基因水平的显著性非常低，才能保证实验水平 I 类错误低；假设发现率控制法在校正 p 值时不那么保守。这种方法是将各独立检验所得到的每个基因的 p 值由小到大依次排列，然后根据其位置得到一个新的界值，方法如下：

基因	g_1	g_2	g_3	…	g_R
p 值	p_1	p_2	p_3	…	p_R
新界值	$\alpha_e(1/R)$	$\alpha_e(2/R)$	$\alpha_e(3/R)$	…	$\alpha_e(R/R)$

其中 $p_1 < p_1 < p_3 < \cdots < p_R$。将所得到的新界值与对应的 p 值比较，那个小于相对应的新界值的 p 值被称之为最大 p 值（LP），被用做分界点。这时当 p 值小于 LP 时，H_0 假设被拒绝，因而被当做差异表达基因。例如当 $\alpha_e(3/R)$ 正好大于 p_3 时，p_3 被选为 LP，因此 g_1，g_2，g_3 的 H_0 被拒绝，就是说在 g_3 右边的那些基因被认为是没有差异显著性。

置换校正法是更通用的 p 值校正法。方法是随机在对照或实验组中选择一些基因，重新计算 p 值。例如表 7-3 中，如果前 5 列是实验组，其余为对照组，通过 t 检验可计算得到各基因实验组和对照组两组比较的 p 值。

表 7-3 基因差异表达 t 检验分析

基因	T_1	T_2	T_3	T_4	T_5	C_1	C_2	C_3	C_4	C_5	p 值
G1	0.800	0.523	0.676	0.806	0.921	0.224	0.696	0.366	0.439	0.460	0.017
G2	0.277	0.609	0.825	0.272	0.224	0.226	0.843	0.124	0.540	0.454	0.983
G3	0.847	0.125	0.529	0.679	0.118	0.815	0.270	0.603	0.037	0.530	0.968
G4	0.408	0.165	0.055	0.307	0.523	0.825	0.792	0.755	0.831	0.928	0.000
G5	0.130	0.524	0.590	0.570	0.695	0.468	0.619	0.721	0.779	0.570	0.279
G6	0.762	0.538	0.356	0.961	0.319	0.165	0.104	0.543	0.430	0.639	0.225
G7	0.468	0.872	0.604	0.289	0.205	0.098	0.004	0.153	0.236	0.974	0.383
G8	0.562	0.939	0.976	0.424	0.948	0.670	0.150	0.529	0.430	0.671	0.100
G9	0.574	0.903	0.020	0.488	0.245	0.924	0.150	0.314	0.094	0.097	0.566
G10	0.047	0.247	0.665	0.430	0.293	0.626	0.827	0.130	0.481	0.418	0.331
G11	0.068	0.679	0.487	0.344	0.756	0.870	0.121	0.636	0.976	0.906	0.271
G12	0.957	0.454	0.730	0.026	0.358	0.389	0.589	0.612	0.107	0.311	0.591
G13	0.369	0.115	0.499	0.244	0.915	0.134	0.393	0.571	0.265	0.312	0.566
G14	0.312	0.364	0.361	0.395	0.319	0.821	0.623	0.796	0.065	0.178	0.383
G15	0.038	0.561	0.842	0.099	0.600	0.986	0.305	0.308	0.083	0.423	0.976

续表 7-3

基因	T_1	T_2	T_3	T_4	T_5	C_1	C_2	C_3	C_4	C_5	p 值
G16	0.342	0.991	0.155	0.407	0.872	0.630	0.947	0.163	0.091	0.528	0.726
G17	0.963	0.359	0.117	0.880	0.520	0.513	0.246	0.410	0.578	0.410	0.441
G18	0.683	0.600	0.987	0.366	0.113	0.948	0.352	0.405	0.905	0.958	0.440
G19	0.099	0.023	0.734	0.047	0.387	0.551	0.996	0.075	0.713	0.275	0.246
G20	1.000	0.665	0.189	0.086	0.157	0.067	0.386	0.258	0.956	0.907	0.714

$T_1 \sim T_5$ 为实验组或处理组；$C_1 \sim C_5$ 为对照组

进行置换分析时，上表中任何一列既可当做实验组，也可当做对照组，但不能同时用于同一个 t 检验。置换校正方法的过程如下：①随机置换上表中的 n 列；②计算得到每个基因的 t 值；③计算每个基因的校正 p 值：第一，得到置换后 P 值小于实际 p 值的置换数目 N；第二，除以置换总数 T 得到校正 p 值 $= N/T$。

值得注意的是，必须采用随机置换大样本的原则才能保证此方法的准确性。一般而言，置换数目需要在 1 000 左右。这就要求最少需要 6 个处理组样本和 6 个对照样本。这种方法比较机械，没有太多理论性可言。

7.6.2 主成分分析

数据集中，一些数据并不能增加有用的信息量，而只是混淆数据，这时需要减少数据的维数(dimension)。微阵列的数据集通常非常大，含有上万个基因，以及很多次不同实验的结果。每个基因以及每个实验就可当做一维。减少维数的方法有多种，主成分分析(PCA)是最常用的一种。

PCA 的目的首先是发现数据集中的相关性变量，然后转换为一组不相关的变量，称之为主成分。这些主成分分别代表线性相关变量中最大变异，即为原始变量的线性组合。每个主成分所表达的信息量取决于其变异。因此，主成分按变异大小依次排列。研究者可以选择其中变异性较高的一种或几种进行进一步分析，而不丧失其原始信息。另外，PCA 不存在相互作用(或相互相关性)，这样，研究者无须考虑原始变量中复杂的相互作用，而着重分析几个不相关的主成分。因 PCA 算法较复杂，这里不列出详细的计算过程，有兴趣的读者可进一步参考相关的统计学教材。

7.6.3 聚类分析

聚类分析的目的是将具有相似特性的对象进行归类，这也可以看做减少维数的方法之一。聚类分析并不是新建立起来的方法，已有许多算法用于表达数据的分析。微阵列实验中聚类的基本概念是将相似表达的基因归成同一种类型，然后观察该类型的生物学意义。聚类分析方法本身不难理解，难的是如何理解其生物学意义。

许多聚类算法并不是直接基于原始数据的考虑，而是利用对象之间的距离(或相似性)阵列进行归类。因此微阵列数据分析中首要的也是最关键的步骤是选择合适的距离来量化已获得基因间生物学上的相似性。有许多种距离测量方法，包括 Euclidean 距离、Euclidean

距离平方、标准化 Euclidean 距离、线性相关距离和 Manhattan 距离等。微阵列数据分析中常用 Euclidean 距离和线性相关距离。

Euclidean 距离独立处理各坐标并尽量使各矢量坐标间的差异最小。对于下列两基因 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$，Euclidean 距离 $d(X,Y)$ 的计算方法是：

$$D_E(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad \text{（公式 7-13）}$$

举个简单的例子来说明 Euclidean 距离的计算方法，如 $X=(3,4)$ 和 $Y=(0,0)$，则

$$D_E(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}=\sqrt{(3-0)^2+(4-0)^2}=5$$

对于同样的两个基因 X 和 Y，线性相关距离的计算方法是

$$D_R(X,Y)=1-R_{x,y} \quad \text{（公式 7-14）}$$

其中 $R_{x,y}$ 是指矢量 X 和 Y（这里是基因 X 和 Y）的 Pearson 相关系数：

$$R_{x,y}=\frac{S_{xy}}{\sqrt{S_x}\sqrt{S_y}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{（公式 7-15）}$$

Pearson 相关系数 $R_{x,y}$ 的数值位于 -1 和 $+1$ 之间，因此线性相关距离总是位于 0 和 2 之间。

以上两种距离计算方法中，有时得到完全不同的结果，试看下面的例子，分析 3 个基因 5 次不同实验的结果：

$$g_1=(2,3,5,6,8)$$
$$g_2=(100,200,500,600,700)$$
$$g_3=(10,8,6,4,2)$$

Euclidean 距离的计算结果分别为：

$$D_E(1,2)=1\,060$$
$$D_E(1,3)=11.4$$
$$D_E(2,3)=1\,063$$

而线性相关距离的计算结果为：

$R_{1,2} = 0.98$　　$D_R(1,2) = 1 - 0.98 = 0.02$

$R_{1,3} = -0.99$　　$D_R(1,3) = 1 - (-0.99) = 1.99$

$R_{2,3} = -0.98$　　$D_R(2,3) = 1 - (-0.98) = 1.98$

以上计算结果中，根据 Euclidean 距离结果，g_1 和 g_3 应归为一类，而线性相关距离结果中，则应将 g_1 和 g_2 归为一类。自然而然的一个问题是，哪种方法最好？或者说什么情况下适合使用哪种方法？尽管已尝试的许多种距离测定方法在聚类分析时都得到比较好的结果，但没有明确的标准来根据不同的分析类型确定距离测定方法。最终还是需要从生物学角度来评估。通常认为，两个基因或两个实验之间的距离取决于所得到的原始数据是相对值（如比值）还是绝对值。但无论是相对值还是绝对值，样品（或实验）聚类中所计算得到的 Euclidean 距离是相同的，而基因聚类的结果则有所不同。线性相关距离则相反，基因聚类的结果相同，但样品聚类的计算结果有所不同。

距离除了用于上述聚类外，它还可以进一步用于聚类算法，如等级聚类（hierarchical clustering）和 K－均数聚类（k－means clustering），即丛间距离（inter－cluster distance）的计算，包括单一连锁（single linkage）、完全连锁（complete linkage）、质心连锁（centroid linkage）和平均连锁（average linkage）等方法。单一连锁，又称最小连锁，是计算不同丛的对象之间的最小距离。完全连锁，又称最大连锁，即计算所比较的丛中成员之间的最大距离。质心连锁计算两丛中心之间的距离。平均连锁是指两丛的各成员的平均距离。

等级聚类是从单丛开始反复合并两个最近的丛或者从整套资料开始反复拆分各丛。前者称为由底到顶或汇聚方法（agglomerative），不断将小丛汇入大丛。后者被称为由顶到底或分割法，将大丛不断拆分为小丛。汇聚方法相对简单，先按两两配对计算各丛间距离，然后将两个距离最近者合并，接着计算其他各丛与新合并丛间的距离，再将距离最近者合并，不断重复上述过程，直到不能将所有的丛被合并到一个大丛为止。其基本过程总结如下。

（1）将每个基因归到自己所在的丛；

（2）找出最近的丛加以合并为一个新丛；

（3）比较新丛与其他丛间的距离；

（4）重复步骤（2）和（3）。

举例来说，某个基因在 6 种不同的实验条件下（D_1，D_2，D_3，D_4，D_5，D_6）分别测得的表达值为 50，100，225，500，700，800。以下用 Euclidean 距离和质心连锁方法演算 $D_1 - D_6$ 的聚类。

首先，最接近的两组数据为 D_1（50）和 D_2（100），因此得到第一个丛，其质心等于 75。

$D_1 - D_2$　　质心 = 75

接着对 75，225，500，700，800 进行聚类，D_5（700）和 D_6（800）最接近，成为第二个丛，其质心为 750。

$D_5 - D_6$　　质心 = 750

现在分析数组 75，225，500，750，最相近的两个值为 75 和 225。得到质心为 150 的丛。

$D_1 - D_2 - D_3$　　质心 = 150

重复上述过程，最后得到图 7－6 所示的等级聚类图。等级聚类分析中常借助 Cluster 3.0 和 Treeview 1.6 等软件进行分析。

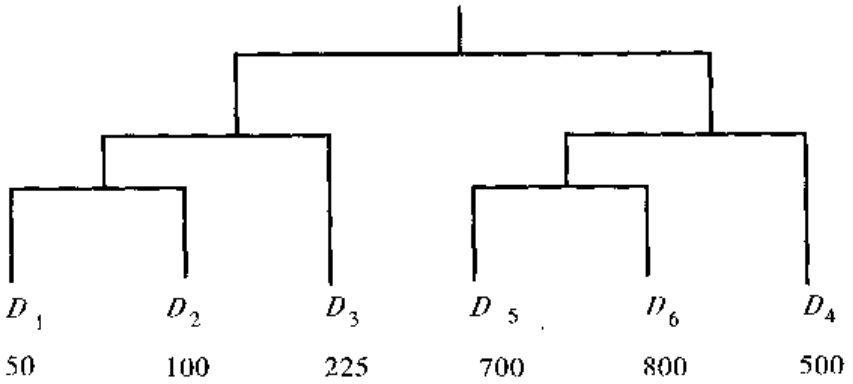


图 7－6　等级聚类分析

基于分配(partition)原理的方法与等级聚类法不同。这种方法试图将数据分成不同的组，如 K－均数聚类法。与等级聚类法不同的是，K－均数聚类法没有预先归好的丛，而是以用户自己定义的 K 丛作为输入参数。接着程序随机选择 K 个点作为各丛的中心，运算各种可能的模式，然后计算出该中心到其他各丛中心的距离。每个模式与最接近的丛相关联。其过程总结如下：

(1)基因被人为地区分为 K 个质心，计算各质心的位置；

(2)每个基因按最小距离被归类到其中某个丛；

(3)重新计算各质心的位置；

(4)重复步骤(2)和(3)。

K－均数聚类法具有一些特点。首先，各模式成员与各丛的关系在后续计算中可发生改变，而不像等级聚类那样保持不变。其次，如果某个质心设定值偏离太远，该丛中可能没有成员。再次，由于初设质心的不同，K－均数法不一定会得到同样的结果。因此，有必要随机选择几组质心，重复上述过程，比较结果是否具有可比性。

在讨论聚类分析的最后，有必要指出的是，不能高估聚类分析的价值。首先，得到一个丛或树状图通常难以回答所涉及的生物学问题。尽管聚类分析可用于类型发现，但通常对于类型比较和类型预测其效力不够。其次，由于基因表达谱涉及数千个基因，所用的算法也有别，测定的距离值也有不同，因此聚类分析是相对主观性的。再次，聚类分析不能提供统计学上有关不同类别中基因表达差异的真实的定量信息。最后，聚类图只是提供一种拓扑结构信息，某个丛中基因的次序和各丛的排列顺序并不代表有用的信息。通常聚类分析只是为某种决策提供支持性的工具，而不是决策性的。

以上所述的等级聚类和 K - 均数聚类分析从机器学习(machine learning)的角度都属于未监测(unsupervised)类型,即其分类预先未知。如果在对不同类型标本进行微阵列分析时,则按已有分类即监测下(supervised)进行判别分析(discriminant analysis 或 classification),这在建立类型预测器时很重要。常用的方法有权重投票(weighted voting)、K - 最近邻居(K - nearest neighbors)、支持性矢量机器(support vector machine, SVM)等。

(胡自华 著 李 巍 译)

参 考 文 献

1 Schena M, Shalon D, Davis RW, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995,270: 467

2 Lockhart DJ, Dong H, Byrne MC, *et al*. Expression monitoring by hybridisation to high - density oligonucleotide arrays. Nat Biotechnol, 1996, 14:1 675

3 Okamoto T, Suzuki T, Yamamoto N. Microarray fabrication with covalent attachment of DNA using bubble jet technology. Nat Biotechnol , 2000,18: 438

4 Schena M, Shalon D, Heller R, *et al*. Parallel human genome analysis: microarray - based expression monitoring of 1 000 genes. Proc Natl AcadSci USA, 1996, 93: 10 614

5 DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 1997, 278: 680

6 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl Acad Sci USA, 2001,98:5 116

7 Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science,1999, 286: 531

8 Alizadeh AA, Eisen MB, Davis RE, *et al*. Distinct types of diffuse large B - cell lymphoma identified by gene expression profiling. Nature, 2000, 403: 503

9 Bittner M, Meltzer P, Chen Y, *et al*. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 2000, 406: 536

10 Lee MLT, Kuo FC, Whitmore GA, *et al*. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA, 2000, 97: 9 834

11 Schuchhardt J, Beule D, Malik A, *et al*. Normalization strategies for cDNA microarrays. Nucleic Acids Res. 2000, 28: E47

12 id kZ. Rectangular confidence regions for the means of multivariante normal distributions. J Am Stat Assoc,1967, 62: 626

13 Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat, 2000, 29: 1 165

第8章 进化遗传学和统计遗传学

进化遗传学(phylogenetics)又称系统发生学,是研究物种进化关系的科学。进化遗传分析就是利用数学、统计方法对这种进化关系进行推断,是生物信息学的重要内容,也是比较基因组学、蛋白组学研究的重要工具。进化遗传学中进化树的分枝状结构(cladistics)与家系中系谱图(pedigree map)的结构相似,所用的运算法则也类似,但后者主要用于遗传方式的确定、连锁分析和遗传咨询。一些与进化遗传分析相关的常用软件如Clustalw,已在第5章描述。本章主要阐述其运算法则和有关原理。

统计遗传学(statistical genetics)是20世纪40年代建立起来的遗传学分支,它涵盖遗传学的常用数理统计方法,特别是遗传流行病学(genetic epidemiology)、群体遗传学(population genetics)、遗传风险估计(risk assessment)中所涉及的计算方法,也是生物信息学的重要内容,已被广泛用于复杂性状的遗传分析、药物基因组学研究、人类学研究和遗传咨询。随着人类基因组计划的完成,统计遗传学在基因组水平有了更进一步的发展,诞生了一门新的分支学科——统计基因组学(statistical genomics)[又称计算基因组学(computational genomics)或基因组生物信息学(genomic bioinformatics)]。广义的统计遗传学或统计基因组学涵盖了微阵列或基因芯片数据处理方法,该领域也逐渐形成了生物信息学的另一重要分支,即微阵列生物信息学(microarray bioinformatics)。相关内容在第7章已作介绍。

8.1 分子进化与遗传距离

8.1.1 进化钟和蛋白质演变速率[1]

人类进化过程中,出现过无数次基因重复和自然选择,其结果是基因家族成员中每个基因的碱基序列发生某些改变。根据所涉及的三联

密码子的改变和兼并性，在蛋白质水平则表现为氨基酸的替换或不改变（沉默突变）。蛋白质进化分析中，通常只考虑氨基酸替换的情况。进化钟（evolutionary clock）是用来量化某种特定蛋白质在进化过程中累积突变的速率，即演变速率（rate of divergence）。演变速率是指每百万年中差异出现的百分数，通常以其倒数来表示，即每出现1%的差别所需的百万年的数目，简称进化期单位（unit of evolutionary period，UEP）。例如通过比对分析，发现珠蛋白在大约85百万年间氨基酸的演变速率为10%，因此，其进化钟表示为0.118%/百万年，或等于8.5UEP。如果人珠蛋白的平均演变速率约为0.096%/百万年（或10.4UEP），现测得β和δ珠蛋白之间氨基酸的差异约为3.7%，那么这两个基因出现演变的时间约在$10.4 \times 3.7 = 39$（百万年）前。

8.1.2 核酸分子的演变和模型

与蛋白质的演变相似，核酸也经历相似的分子进化过程。但只涉及4种核苷酸之间的改变，相对20种氨基酸之间的变化性要少。在第5章，我们讨论了DNA和RNA序列分析的方法，了解到不同种系间DNA和RNA序列的异同是我们构建进化树的依据之一。在分子进化研究中，我们往往首先假定各序列是同源的，它们皆具有单一祖先序列。这一祖先序列在进化过程中发生了一系列的核苷酸突变，从而导致各序列之间出现差异，这种差异在分子进化研究中以遗传距离表示。表8-1表示了各种核苷酸的演变情况。在以上假设的基础上，Judes和Cautor进一步假设每一碱基突变为另外3种碱基中的任何一种的几率相同，其频率常数为$\mu/3$，μ为碱基替换频率。Kimura[2]却认为转换（transition，两种嘧啶或两种嘌呤碱基之间的突变）和颠换（transversion，一个嘧啶和一个嘌呤碱基之间的突变）具有不同的频率，即α和β。表8-2简要说明了以上两种遗传模型。

表8-1 同源序列间的核苷酸置换

祖先序列	同源序列1	同源序列2	演变性质
A	A	A	
C	C	C > A	单一置换（single substitution）
T	T	T	
G	G	G	
A	A > C > T	A	多重置换（multiple substitution）
A	A	A	
C	C > G	C > A	同义置换（coincidental substitution）
G	G	G	
T	T > A	T > A	平行置换（parallel substitution）
A	A	A	
A	A > C > T	A > T	趋同置换（convergent substitution）
C	C	C	
G	G	G	
C	C	C > T > C	回复置换（back substitution）

表 8-2 Judes-Cantor 单参数模型(右上三角)和 Kimura 两参数模型(左下三角)

	A	T	G	C
A		α	α	α
T	β		α	α
G	α	β		α
C	β	α	β	

α、β 分别为转换和颠换的频率。

8.1.3 遗传距离的计算方法

根据以上遗传模型,Judes 和 Cantor 提出了 DNA 序列距离 K(最早为氨基酸序列延伸而来)计算公式:

$$K = \frac{3}{4}\ln\left(\frac{4}{4q-1}\right) \approx 2\mu t \qquad (公式 8-1)$$

其中 q 为同源 DNA 序列中具有相同碱基的概率,经过 t 世代,由于祖先序列的趋异变化,其值为:

$$q_t = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{8\mu}{3}\right)^t \qquad (公式 8-2)$$

μ 为碱基替换频率。

距离 K 适用于显示两条序列从一个祖先序列趋异进化以来的时间,并能用于序列间系统树的构建。在计算时,均需要将序列作初步的列线分析。Kimura 在其两参数模型下证实,由于趋异变化,由转换造成差异(Ⅰ型变化,$P_Ⅰ$)或由颠换造成差异(Ⅱ型变化,$P_Ⅱ$)的碱基,随时间而变化:

$$P_Ⅰ = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)\cdot t} + e^{-8\beta\cdot t})$$

$$P_Ⅱ = \frac{1}{4}(1 - e^{-8\beta\cdot t}) \qquad (公式 8-3)$$

如果 $k = \alpha + 2\beta$ 是单位时间碱基替换的总频率,则适合作为系统树的距离尺度为:

$$K = -\frac{1}{2}\ln[(1 - 2P_Ⅰ - P_Ⅱ)\sqrt{1 - 2P_Ⅱ}] \approx 2kt \qquad (公式 8-4)$$

该类距离可用于有关系统树距离矩阵中。用样本比值代入公式 8-4 就可估计这些距离。

DNA 序列距离 K 又称为 DNA 序列间的演变度,是序列间相异性的一个指标。蛋白质

序列的演变度分为两序列同义变化的演变度(KS)和非同义变化的演变度(KA),根据 Jukes – Cautor 单参数模型和 Kimura 两参数模型等遗传模型,可以分别计算得到两序列的演变度(或称为蛋白质序列间的距离)。

8.2 进化遗传模型

进化遗传学研究的是进化关系,进化遗传分析就是要推断或者评估这些进化关系。通过进化遗传分析所推断出来的进化关系一般用分枝图表(称为遗传树,又称进化树)来描述,包括了分子进化(基因树)、物种进化以及分子进化和物种进化的综合 3 种类型。在现代进化遗传学研究中,研究的重点已经不再是生物的形态学特征或者其他特性,而是生物大分子尤其是蛋白质或核酸序列的进化特征。由于进化的发生是已经完成的历史,因而无法再现,只能依靠建立的数据模型去模拟真实的进化过程,然后假定现有序列之间的差异是基于模拟的进化结果。一个正确的进化模型应能反映真实的进化过程,并有助于我们绘制出正确的遗传树。

进化遗传的建树方法都会预先假定一个进化模型。比如,目前广泛使用的方法都假定进化的演变是分枝型的,因此我们可以用树状拓扑发生图来描述相关数据间的演变关系。但是在一个给定的数据组中,因为存在着物种的杂交以及物种之间遗传物质的传递,这个假定很可能会被推翻。因此,如果所观察的序列并非是严格遵循物种内的遗传规律的话,所用的进化遗传方法就会得到错误的结果。进化遗传分析中固有的模型需要以下的假定:①序列有指定的来源并且正确无误[3]。②序列是同源的(也就是说,所有的序列都起源于同一祖先序列);这些序列不是同系物(paralog)的混合物。序列比对中,不同序列的同一个位点都是同源的。③在接受分析的一个序列组中,序列之间的进化遗传史是相同的(比如,序列中不存在细胞核序列与细胞器序列的混合)。④序列中的所有位点的进化都是随机的。序列中的每一个位点的进化都是独立的。⑤样本序列之间的差异包含了足以解决感兴趣问题的进化遗传信号。

8.3 遗传树的构建

根据 DNA(或 RNA 或蛋白质)的序列或结构差异关系可构建遗传树,又称分子进化树(evolutionary tree)或进化遗传树(phylogenetic tree)。进化树给出分支层次或拓扑图形,每个节点(node)是产生新的基因复制或享有共同祖先的生物体的演变点的一种反映。树枝的长度表示当这些事件发生时就存在的 DNA 序列与后出现的 DNA 序列之间的进化距离。根据进化树不仅可以研究从单细胞有机体到多细胞有机体的生物进化过程,而且可以粗略估计现存的各类种属生物的演变时间。通过 DNA 的分子进化树分析,为从分子水平研究物种进化提供了新的手段,可以比较精确地确定某物种的进化地位。对于物种分类问题,DNA 的分子进化树亦可作为一个重要的依据。

遗传树分有根(rooted)和无根(unrooted)树。图 8-1 中显示 4 个物种 A,B,C,D 两种有根树和一种无根树的表示方式。有根树反映了树上物种或基因的时间顺序,而无根树只反映分类单元之间的距离而不涉及谁是谁的祖先问题。用于构建遗传树的数据有两种类型:一种是特征数据(character data),它提供了单个基因、个体、群体或物种的信息;二是距离数据(distance data)或相似性数据(similarity data),它涉及的则是成对基因、个体、群体或物种的信息,参见 8.1.3 节。距离数据可由特征数据计算获得,但反过来则不行。这些数据可以矩阵的形式表达。距离矩阵(distance matrix)是在计算得到的距离数据基础上获得的。距离的计算总体上是要依据一定的遗传模型,并能够表示出两个分类单位间的变化量。系统树的构建质量则依赖于距离估算的准确性。

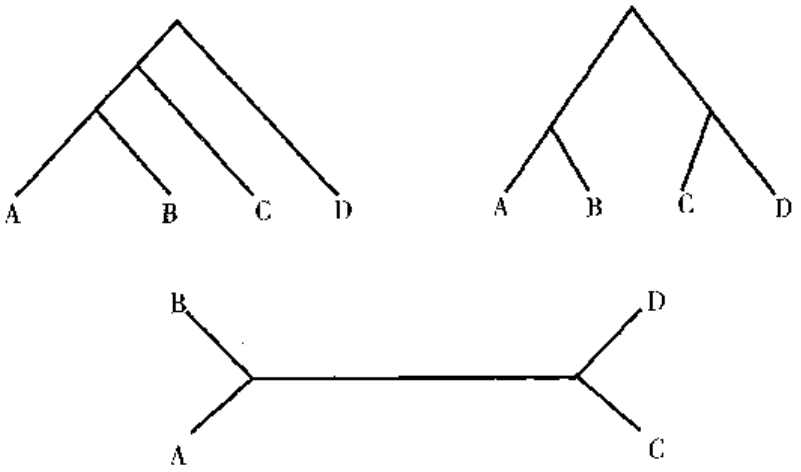


图 8-1 4 个物种(A、B、C 和 D)的 2 种有根树和 1 种无根树形式

在此,我们主要讨论构建序列进化树的一般步骤,包括建立数据模型(序列比对)、建立取代模型、建立进化树以及进行进化树评估等。

8.3.1 建立数据模型

建立一个比对模型的基本步骤包括:选择合适的比对程序;然后从比对结果中提取进化遗传的数据集。至于如何提取有效数据,取决于所选择的建树程序如何处理容易引起演变的比对区域和插入/缺失序列(即所谓的 indel 状态或者空位状态)。

一个典型的比对过程包括:首先应用 Clustalw 程序,然后进行手工比对,最后提交给一个建树程序。这个过程有如下特征:①部分依赖于计算机(也就是说,需要手工调整);②需要一个先验的进化遗传标准(即需要一个前导树);③使用先验评估方法和动态评估方法(推荐)对比对参数进行评估;④对基本结构(序列)进行比对(对于亲水氨基酸,推荐引入部分二级结构特征);⑤应用非统计数学方法优化。这些特征选项的取舍依赖于进化遗传分析方法。有关比对方法参见第 5 章。

8.3.2 决定取代模型

取代模型既影响比对,也影响建树,因此需要采用递归方法。对于核苷酸数据而言,可

以通过取代模型中的两个要素进行计算机评估，但是对于氨基酸和密码子数据而言，没有什么评估方案。其中一个要素是碱基之间相互取代的模型；另外一个要素是序列中不同位点的所有取代的相对速率。还没有一种简单的计算机程序可以对较复杂的变量（比如，位点特异性或者系统特异性取代模型）进行评估，同样，现有的建树软件也不可能理解这些复杂变量。

8.3.3 建树方法

3种主要的建树方法分别是遗传距离建树法、最大似然法（maximum likelihood，ML）和最大节约法（maximum parsimony，MP）。

8.3.3.1 基于遗传距离的建树方法

距离树考察数据组中所有序列的两两比对结果，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度。距离方阵用来简单地计算两个序列的差异数量值。这个数量值被看做进化距离。其准确性大小依赖于进化模型的选择。然后运行一个聚类算法，从最相似（也就是说，两者之间的距离最短）的序列开始，通过距离值方阵计算出实际的进化树，或者通过将总的树枝长度最小化而优化出进化树。

距离建树方法根据双重序列比对的差异程度（距离）建立进化树。如果所有的进化演变事件都很精确地记录在序列中，那么距离建树方法将会重构真实的进化树。然而，当序列突变达到饱和时，演变程度就会达到上限。一对演变序列之一在某个特定位点发生突变后，后续的突变不管发生在哪一条序列，都不会再造成更多的差异。实际上，后续的变异很可能会使得前一个位点恢复成相同的状态，从而屏蔽掉前一个突变，因此，绝大多数基于距离的建树方法都会针对这种“不可见”的替换进行修正。在实际操作中，使用的速率方阵可有效地假定：在所观察的相同的碱基对中，实际上存在一定比例的位点，这些位点经受了多次突变，而且，随着整个序列差异程度的增加，这些位点的比例也在增加。有些程序（至少是随意地）会计算出不正确的距离，比方说，MEGA程序[4]只针对密码子和氨基酸数据会计算出错误的距离；除非序列的整体差异程度很小，事实上，这个程序肯定会得到错误的计算结果。

双重序列差异使用最大似然法的取代速率来计算。最常用的距离建树程序使用的模型局限于时间可逆模型，只包含很有限的几个取代模型；但是PAUP 4.0基本上能测试时间可逆模型的所有变化，包括用最大似然方法从数据中评估得到的实际模型，以及针对非静态数据的距离建树方法所得到的模型。

与最大似然法相比，距离建树方法的计算强度很小，但是可以使用序列进化的相同模型，这是其最主要的优点。该法的缺点是屏蔽了真实的特征数据。最常用的距离建树方法是“不加权配对组算术均数”（UPGMA，unweighted pair group method with arithmetic mean）、相邻连接方法（NJ，neighbor joining）和对距离进化树的附加值进行优化的方法，包括最小进化方法（ME，minimum evolution）。有些方法被多个软件包引用，但是在使用时不一定会使用相同的指定参数和（或）进化树优化特征。

（1）不加权配对组算术方法（UPGMA） UPGMA是一种聚类或者说是分类方法。它按照配对序列的最大相似性和连接配对的平均值将进化树的树枝连接起来。它不是一种严格

的进化距离建树方法。只有当序列演变是基于一个分子钟或者近似等于原始的序列差异性的时候，我们才会期望 UPGMA 产生一个拥有真实的树枝长度的准确的拓扑结构。正如前面讨论的那样，在实际应用中很少会遇到这种情况。

（2）相邻连接法（NJ） 相邻连接算法在距离建树中经常会用到，而不在乎使用什么样的优化标准。完全解析的进化树是通过对完全没有解析的“星型”进化树进行“分解”得到的。分解的步骤是连续不断地在最接近（实际上是最孤立的）的序列对中插入树枝，而保留进化树的终端。于是，最接近的序列对被强化了，而“星型”进化树也被改善了，这个过程将不断重复。这种方法运算很快，也就是说，对于一个 50 个序列的进化树，只需要若干秒甚至更短时间。

（3）Fitch – Margoliash 算法（FM） FM 法设法将所有可能观察到的距离相对于进化树中所有可能的路径长度的偏差的平方极小化，将进化树中观察到的双重距离的合适度极大化[5]。有一些变量，在如何对误差进行加权的处理上不太一样。对变化的估计值并非是完全独立的，因为所有的进化树内部的树枝中存在的误差都至少被计算了两次[6]。

（4）最小进化方法（ME） 最小进化方法先使用与 FM 相同的方式计算出路径长度，然后根据路径长度优化出最短的进化树；也就是说，它要求将观察到的距离相对于基于进化树距离的偏差平方最小化[5,6]。与 FM 法不同的是，ME 方法并不使用所有可能的双重序列距离和所有可能的相关的进化树路径长度，而是先根据到外层节点的距离固定进化树内部节点的位置，然后根据这些观察点之间的最小计算误差，对内部的树枝长度进行优化。因此，这种方法可能消除了 FM 计算的依赖性。

（5）距离建树软件的评价 ME 和 FM 似乎是最好的程序，它们在模拟研究中的得分几乎相同[7]。ME 在计算机程序中的应用越来越广泛，包括 METREE[6] 和 PAUP 都在使用 ME。对于蛋白质数据，PHYLIP 中的 FM 程序提供了最多的时间可逆取代模型，但是没有对位点内部的取代速率差异进行修正。MEGA 和 METREE 软件包含有一个针对蛋白质的修正，这种方法只有当演变度很小时才合理。

模拟研究指出，对于一个大范围的进化树形状空间，UPGMA 的可操作性很差。我们并不赞成使用这种方法，而之所以在这里会提及这个方法，是因为目前的出版物上经常会出现 UPGMA 的应用，这一点可以由当前的出版物中出现的 UPGMA“基因进化树”证明[7]。

很显然，NJ 是最快的程序，并且所产生的进化树同 ME 进化树相比，虽不能说一样，但也已经非常相近了[6]。但是，NJ 只产生一个进化树。根据数据结构，有大量不同的进化树可能和 NJ 进化树一样好，甚至比 NJ 进化树好得多。

8.3.3.2 最大似然性建树方法

最大似然方法考察数据组中序列的多重比对结果，优化出拥有一定拓扑结构和树枝长度的进化树，这个进化树能够以最大的概率导致考察的多重比对结果。最大似然法系统树的构建试图避免其他方法的局限性，尽管它需要的计算量大得惊人。与距离矩阵法不同，似然法充分有效地利用所有资料而不是将资料简缩为距离的集合。它们与简缩法的不同之处在于其进化概率模型采用了标准的统计方法[8]。当考虑实施最大似然法时，该方法先假定系统树的形式，然后选择分枝长度以使产生特定系统树的资料的似然值最大化。通过比较不同系统树的似然函数值，将具有最大似然值的系统树看做最佳估计。一个直接的问题是

随着 OUT 的增加，系统树的数目迅速增加。当树端具有 n 个 OUT 时，无根树的数目为 $(2n-5)!/[2(n-3)!]$。具有 n 个树端的有根树数目与具有 $n+1$ 个树端的无根树数目相同。实际应用时，只研究所有系统树的一个亚集。对于 DNA 序列资料，似然法依据的模型规定了在特定时间内由于突变使一个序列变更为另一序列的概率。尽管 DNA 序列中的毗邻碱基不是独立的，但是模型的确假定了不同位点上进化的独立性，从而某系统树上一组序列的概率就是序列上每一位点概率的乘积。

8.3.3.3 最大节约法建树方法

最大节约方法考察数据组中序列的多重比对结果，优化出的进化树能够利用最少的离散步骤去解释多重比对中的碱基差异。用最大节约方法搜索进化树的原理是要求用最小的改变来解释所要研究的分类群之间观察到的差异。

8.3.4 遗传树搜索

单一的遗传树的数量会随着分类群数量的增长而呈指数增长，从而变为一个天文数字。由于计算能力的限制，现在一般只允许对很小一部分的可能的遗传树进行搜索。具体的数目主要依赖于分类群的数量、优化标准、参数设定、数据结构、计算机硬件以及计算机软件。有两种搜索方法可保证找到最优化的遗传树：穷举法和树枝跳跃法(BB)。对于一个很大的数据集，这两种方法都很不实用。对分类群数量的限制主要取决于数据结构和计算机速度，但是对于超过 20 个分类群的数据集，BB 方法很少会得到应用。穷举法要根据优化标准，对每一个可能的进化树进行评估。BB 方法提供一个逻辑方法，以确定那些进化树值得评估，而另一些进化树可被简单屏蔽。因此 BB 方法通常要比穷举法快得多。

绝大多数分析方法都使用“启发式”的搜索。启发式先搜索出相近的次优化的进化树家族(“岛屿”)，然后从中得到优化(“山顶”)。不同的算法用不同程度的精确性搜索这些岛屿和山顶。最彻底也是最慢的程序(TBR，tree bisection - reconnection，进化树对分重接)是先把进化树在每一个内部树枝处劈开，然后以任意方式将劈开的碎片重新组合起来。最快的算法只是检查一下相邻终端的不太重要的重新组合，因此倾向于找到最近的岛屿的山顶。

降低搜索代价的最好方法是对数据集进行剪除。影响优化搜索策略选择的因素(数据量、数据结构、时间量、硬件、分析目的)太复杂，无法推荐一个简单可行的方案。因此进行搜索的用户必须对数据非常熟悉且有明确的目标，了解各种各样的搜索程序及自己的硬件设备和软件的能力。

除上述当前应用最广的方法外，还有大量的建立和搜索进化树的其他方法。这些方法包括：Wagner 距离方法和亲近方法(距离转化方法)；Lake 的不变式方法(一个基于特征符的方法，它选择的拓扑结构包含一个意义重大的正数以支持颠换)；Hadamard 结合方法(一个精细的代数方阵方法，对距离数据或者观察到的特征符进行修正)；裂解方法(这个方法决定在数据中应该支持哪一个基于距离的可选的拓扑结构)；四重奏迷惑(Quartet puzzling)方法(可以为最大似然性建树方法所应用，这个算法相对而言是个较快的进化树搜索算法)。

8.3.5 确定树根

上述的建树方法所产生的都是无根树(进化树没有进化的极性)。为了评估进化假说,通常必须要确定进化树的树根。确定进化树的树根并不是一个简单的问题。一种确定树根的好方法就是分析时加入一个复制的基因。如果来自绝大多数物种或者所有物种的所有的平行基因在分析时都被包含进去,那么从逻辑上我们就可以把进化树的树根定位于平行基因进化树的交会处,当然要假定在所有进化树中都没有长树枝问题。

8.3.6 评估进化树

现在已经有一些程序可以用来评估数据中的进化遗传信号和进化树的健壮性。对于前者,最流行的方法是用数据信号和随机数据作对比实验(偏斜和排列实验);对于后者,可以对观察到的数据重新取样,进行进化树的支持实验(非参数自引导和对折方法)。似然比例实验可以对取代模型和进化树都进行评估。

8.4 MCMC 方法在进化遗传分析中的应用

马尔可夫链蒙特卡罗分析方法(Markov Chain Monte Carlo,MCMC)是一种重要的模拟计算方法,MCMC 的研究对建立可实际应用的统计模型开辟了广阔的前景。20 世纪 90 年代以来,很多应用问题都存在着分析对象比较复杂与正确识别模型结构的困难。现在根据 MCMC 理论,通过使用专用统计软件进行 MCMC 模拟,可解决许多复杂性问题。此外,得益于 MCMC 理论的运用,使得贝叶斯(Bayes)统计得到了再度复兴,以往被认为不可能实施计算的统计方法变得很轻而易举了。以前,MCMC 主要在物理学中应用,而现在大量生物信息学问题的解决都和 MCMC 有关。

8.4.1 贝叶斯定理

运用贝叶斯定理重建遗传树时,要求我们首先定义一个参数,此参数包含所有可能的组合状态,$\phi = \{T_i, V_i, \theta\}$,$T_i$代表第 i^{th}个可能的树的构象,V_i 代表与这个构象相关联的树枝的长度,而 θ 则是一组可用于我们评估模型的参数。根据贝叶斯定理,我们可以计算当给定序列数据 X 时,组合状态后续的可能性,即

$$p(\phi \mid X) = \frac{P(X \mid \phi)p(\phi)}{\sum_{i=1}^{N_s} \int dv_i \int d\theta P(X \mid \phi)p(\phi)} \qquad \text{(公式 8-5)}$$

其中,N_s代表当给定 s 个种属的数据时,可能的树形个数;$P(X|\phi)$ 是这些数据的可能性;$p(\phi)$ 是和状态 θ 相关的先验的可能性。

8.4.2 MCMC 的基本概念

对一个实际规模的问题，按公式 8－5 计算其分母是不可能的，因而 MCMC 便得以应用。一个标准的广义快速 MCMC 方法是通过重复性两步法，在我们的状态空间 Ω 上构建一个马尔可夫链。首先，从实际的状态 ϕ_n 推导出一个新的状态 ϕ'，然后依据这两个状态 ϕ_n 和 ϕ' 及其推论的后续发生的可能概率，这个推导的状态可以被接受或拒绝。在非常微弱的情况下，这条链可以达到一个均衡的状态。通过抛弃这条链的最初部分，可以依据后续的可能密度 $p(\phi|X)$，推导出其状态。现在，已有一个基于 MCMC 的原理，叫 PHASE 的程序，可以从后续的可能密度产生大量的数据，然后从这些数据计算出任何感兴趣的同一进化遗传特性的后续可能性。比如对于一个特定的树的结构的后续可能性可以很方便地以它在 PHASE 产生的 MCMC 的数据中出现的频率来表示。这样，我们也可以使后续可能的密度曲线与 γ 分布的参数相吻合。

8.4.3 MCMC 的建树方法

树的结构每隔 10 个循环，就依据最近的邻居交换(NNI)的方法(图 8－2)[9]或次级树的修剪(SPR)方法(图 8－3)进行重建。每个循环，随机挑取的树枝长度被从正态分布的中心得到的数值 δ 修改。当树枝的长度变为负值时，特殊的规则将被应用，以使树的结构得以改变[9]。如果树枝是内部树枝，最近的两个树枝之一将被交换，以使它们具有相同的可能性。这就是上述的最近邻居交换(NNI)法。新的内部树枝的长度被设定为 $y = |x + \delta|$。假如树枝是末端树枝，NNI 方法就不能应用，而代之应用其镜像。对 SPR 和 NNI 的方法而言，其接受的比率通常都较低。不过经树枝长度修改引导的“本地”的 NNI 方法，有相对较好的接受度。

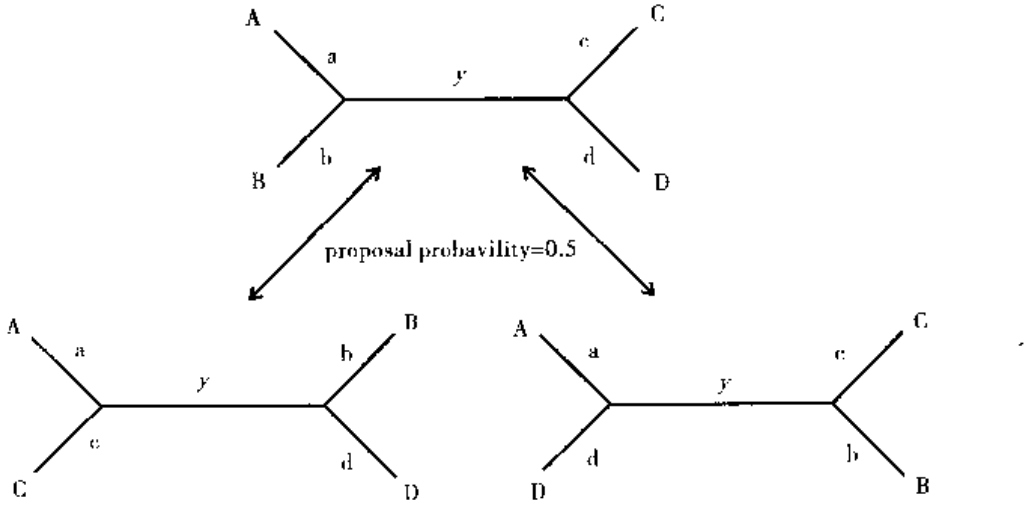


图 8－2 最近邻居交换(NNI)的方法

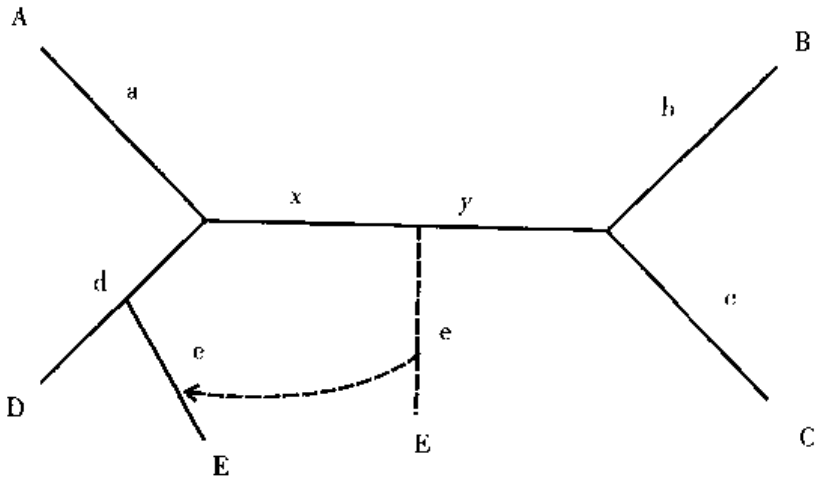


图 8-3 次级树的修剪(SPR)方法

8.4.4 MCMC 的难点

我们所采集到的生物数据是有限的,但是真实的生物进化历程却是未知的,因此,对进化树的人为估计、计算极易带来不确定性。这种不确定性不但体现在树的枝长,而且主要体现在树的拓扑结构上。我们可以怀疑是否最大似然法真的总是发现真实的最大似然功能。同样,应用 MCMC 方法时,马尔可夫链也会无法计算出稳定分布的后续可能性。造成这种结果的其中一种原因就是由于似然性曲线的局部最大化无法访问所有的参数区间的高概率区域。不过,不成熟的方法本身和(或)没有运行此链达到足够的长度是导致数据缺陷的主要原因[10]。遗憾的是,并不是总是很容易地辨别出是何种原因造成这种缺陷。我们的建议是运行此链达到足够的长度,监控几个模型参数的产生情况(因为只监控似然性是不够的),使用随机的树的起始位点重复同一实验,以检验是否所有的链给予相似的结果(比如,替代模型参数,一致性,似然性等)。

(汤 斌)

8.5 疾病基因的连锁与关联分析[11~16]

疾病基因的克隆是目前遗传学研究的热点之一,特别是有关多基因性状或数量性状的定位与克隆。对这种疾病基因的分析,关键的 3 个步骤是收集样本资料(特别是系谱资料)、个体的基因型分析以及数据资料处理。其中所涉及的统计遗传学方法分为连锁分析(linkage study)和关联分析(association study)两大类。

连锁分析是建立在系谱的基础上,主要通过计算重组率,来反映两基因位点(其中之一为疾病基因位点)之间的近邻关系,即通过检验假设和统计检验,推断两位点是否存在连锁。连锁分析的常用数理统计原理有两种,一是贝叶斯分析(Bayesian analysis),二是似然性比值法(likelihood ratio test, LRT),后者常用对数优势记分法(LOD score)表示,简称 Lods

法。对于大样本的连锁分析，也可用正态分布的 Z 检验。

关联分析一般是基于群体的(population - based)病例对照研究(case - control study)的实验设计而进行，分析某种基因型或单体型与疾病的关联性，这是遗传流行病学的基本方法。统计方法多用相对危险度(RR)或几率比值(OR)，参见 10.4 节。实际应用中，对于某些具有家族聚集现象的病例对照资料的关联分析，常用非参数统计检验，主要包括两种：①传递不平衡分析(transmission disequilibrium test, TDT)(统计软件常用 GENEHUNTER：http://www.fhcrc.org/labs/kruglyak/downloads/index.html)；②受累同胞对分析(affected sib - pair analysis, ASP)(统计软件常用 ASPEX：ftp://lahmed.stanford.edu/pub/aspex/index.htm)。TDT 分析是比较某些特殊等位基因由亲代向子代的传递频率的差别。这种方法的主要受限制因素是必须获得患者父母的 DNA 样品，特别是那些发病年龄较晚的疾病如早老性痴呆症、肿瘤、II 型糖尿病等，往往难以获得患者双亲的 DNA。这通常需要采用修正的 TDT 方法进行分析。另外，常用的补救方法是仅分析患者及其同胞的等位基因频率，即 ASP 分析。特别在研究高发病率的性状时，ASP 方法的检验效能要优于 TDT 法。

数量性状是多基因遗传的基础。在 2.3.2 节中已介绍过有关多基因遗传方式及其特点。有关数量性状位点(quantity trait loci, QTL)的定位是目前人类遗传学的研究热点之一。鉴定多基因病的相关基因或易感基因的常用方法有两种：一是候选基因研究，二是在全基因组扫描基础上的定位克隆，参见 3.5 节。由于一些多基因病的易感基因与邻近位点存在共分离现象(cosegregation)，即连锁不平衡现象(linkage disequilibrium, LD)，通过连锁分析可计算出邻近位点与易感基因间的遗传距离，反映其连锁关系，用于基因的定位和克隆；也可通过基于 LD 基础上的单倍型分析该邻近位点与疾病的关联性，来确定易感基因位点，又称连锁不平衡作图分析(详见 9.2.2 节)。

因此，在进行与疾病基因定位(特别是多基因病易感基因)分析之前，必须对研究资料进行分类，再决定所采用的统计分析方法。分类方法通常基于以下几点。

(1)样品类型　根据样品来自家族成员还是无亲缘关系人员区分为家族分析和群体分析两大类。家系分析中依据亲代数目和同胞数目进行细分，采用的实验设计和统计方法有所差别。

(2)标记物种类　传统的关联分析方法是分析具有一对等位基因的单一分子标记，如果属于多等位基因或多个分子标记物，或利用单体型分析，则采用的统计方法也有区别。

(3)其他因素　鉴于是否存在基因间相互作用或基因与环境相互作用、统计方法的分析效能大小、是否具有可靠的分析软件等综合考虑实验设计和统计检验(常用统计软件目录参见http://linkage.rockefeller.edu/soft/list.html)。有关复杂性状易感基因研究的实验设计和统计方法的选用参见图 8 - 4。

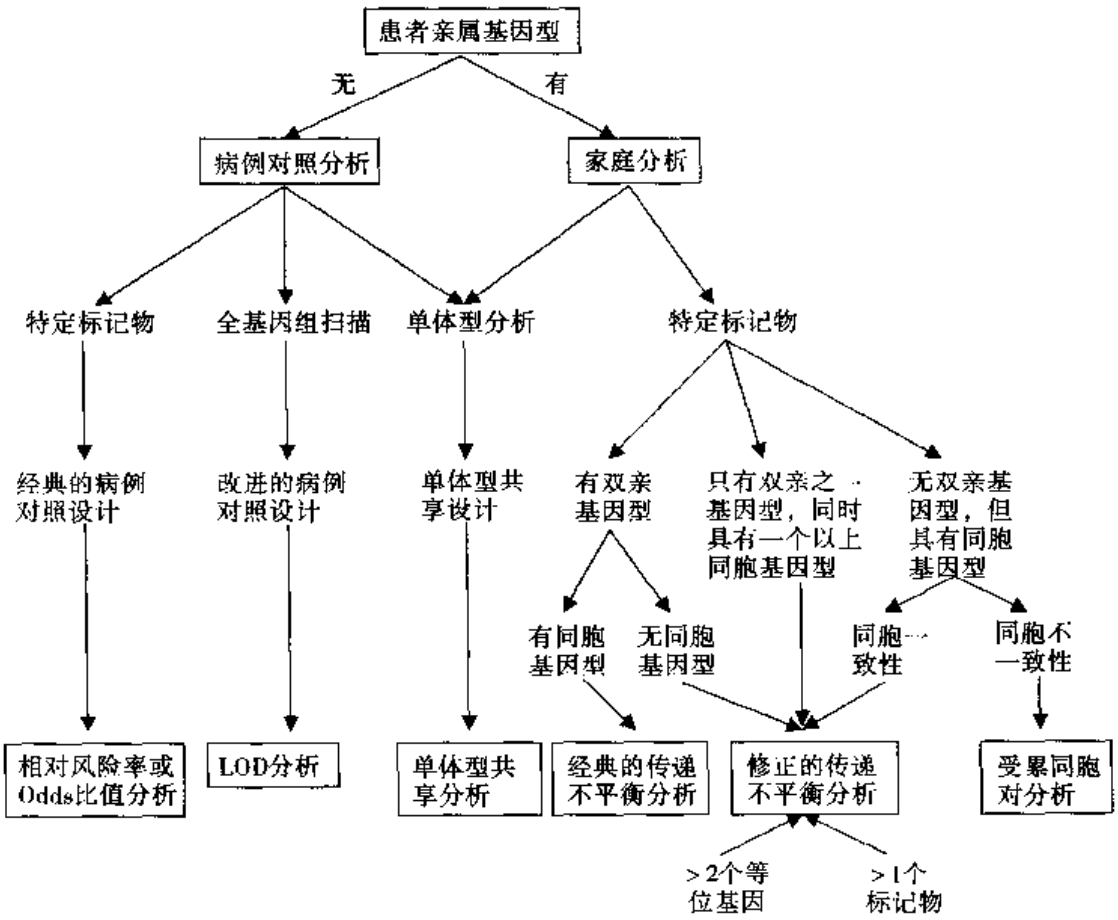


图 8-4 复杂性状易感基因的定位分类方法

8.5.1 家族聚集性分析

通过临床观察，可得知某种疾病是否具有家族聚集性(familial aggregation)，但要证实这种家族聚集性，必须依靠统计学方法。通常采用的方法有：①比较患者亲属与普通人群的患病率(prevalence rate)或发病率(incidence rate)的差别；②比较有家族史患者亲属与无家族史患者亲属的患病率或发病率的差别；③比较不同亲缘级别的患者亲属的患病率或发病率的差别；④比较某种数量性状在患者同胞对之间的相关性和在非同胞对之间的相关性的差别。

这种家族聚集性的存在，通常表明患者家族中存在某种致病基因，或者患者家族暴露于同一危险因素的环境中。要回答家族聚集性是否为遗传性，最常用的分析方法为双生子分析法(twin study)，双生子可区分为单合子双生子(monozygote, MZ)和双合子双生子(dizygote, DZ)。双生子分析法中，主要比较 MZ 和 DZ 中疾病一致性和不一致性的配对数目，计算各自的频数分布，如果某性状在 MZ 中出现的比例高于 DZ 的比例，并具有统计学意义，则表示该性状与遗传因素有关。

要回答遗传与环境的相互作用，需要利用生态遗传学(ecogenetics)的分析方法。通过

研究分居(不同环境)和同居(相同环境)MZ 中的性状分布规律,可以揭示遗传或环境因素的关系。另外,收养研究(adoption study)也是研究性状与遗传或环境因素是否相关的常用方法。此方法中,是比较具有相同后天环境下的收养者非生物学亲属中某性状的重现率,或者比较不同后天环境下收养者生物学亲属中该性状的重现率。由于社会因素的影响,后一种比较方法通常难以实现,实验设计中主要是利用前一种比较方法。如果收养者(本身为患者)非生物学亲属中某种疾病的患病率与非收养的患者亲属中的患病率相同,说明遗传因素起一定作用;如果两者差别有统计学意义,说明环境因素起一定作用。收养研究设计的一种改进方法是利用半同胞对设计,其分析原理基本相同,这在离婚率高的社会中尤其适用。

值得注意的是,家族聚集性分析中须特别注意排除实验设计中混杂因素的影响,如年龄、性别、种族等。

关于复杂性状或数量性状的定位,是目前遗传学研究中的热点。单基因病研究中的分离分析方法大大促进了孟德尔遗传病基因的定位克隆。这种分析方法对于某些符合孟德尔遗传的肿瘤综合征(多为常染色体显性遗传)也适用,例如家族性乳腺癌中 BRCA1 和 BRCA2 基因的克隆。这就是说,对于某些多基因病,通过家族聚集性分析,可细分为不同的亚类,通过类似单基因病的分离分析方法,可逐个解析不同基因的贡献。但对于大多数多基因病而言,要找到这种具有明确孟德尔遗传特征的亚类是很困难的。这种多基因病或复杂性状的分析主要依靠下文将要详细介绍的连锁分析和关联分析方法。这两种方法也常建立在对患者家庭成员分析的基础上。

8.5.2 LOD 分析

连锁分析主要通过分析致病基因或性状基因与已知标记基因是否存在连锁关系,用于待分析基因的染色体定位。统计分析方法包括经典连锁分析和非参数连锁分析两种常用方法。非参数连锁分析方法通常用于基因的遗传模式不明确的情况(即不知道分布类型),只需要知道双亲和子女各位点的基因型,无须考虑基因型频率。这在下文 TDT 和 ASP 分析中详细叙述。

经典连锁分析中,其前提是标记基因和目的基因均符合经典的孟德尔遗传规律。其分析方法主要为系谱分析和最大似然法(maximum likelihood)。系谱分析法最直观,是临床遗传学和遗传咨询中最常用的方法。连锁分析根据位点数目分为单位点、双位点和多位点模型。其中以双位点模型最常用,特别适合于疾病基因的定位,其最大似然法概率的计算公式为:

$$P_r = \frac{\prod_{i=1}^{n} p_i(\theta)}{\prod_{i=1}^{n} p_i(0.5)} \qquad \text{(公式 8-6)}$$

式中,P_r表示连锁相对概率;P_i表示第 i 个家系在重组率为 θ 或 0.5(不连锁)时的概率;n 为家系数,Π 表示各计算概率的乘积。P_i的计算根据二项分布(binomial)函数而展开。公式 8

-6 中 P_i 的计算通常是根据分子和分母取最大似然性函数，再得到其比值（likelihood ratio，LR），此法又称为似然性比值检验（LRT），或称 G 检验。实际应用中，常以 LR 的常用对数 lg(LR) 来表示连锁相对概率，称之为对数优势记分法（LOD score），简称 Lods 值。Lods 值的计算常用现行软件，如 GENEHUNTER，LINKAGE（http://linkage.rockefeller.edu/soft）等给出。当 Lods 值 >1，提示有连锁关系；Lods 值≥3，表示肯定连锁；Lods 值≤ -2，否定连锁关系；当 Lods 值介于 -2 和 +1 之间时，需要增大样品量，直到 Lods 值大于 1 或小于 -2。

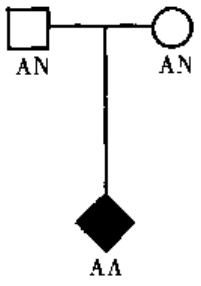
连锁分析的单位点模型主要是重组率 θ 的计算，θ 是用来表示重组型（发生交换）配子占总配子的百分比。通常将重组率为 1%［即（=0.01）］时的两个位点之间的距离称为 1 厘摩（centimorgan，cM）；当重组率为 50%，即图距为 50 cM，θ = 0.5 时，认为不连锁。1 cM 的物理距离平均约为 1 Mb。重组数目（r）与遗传距离（d）的相互关系根据 1919 年 Haldane 方程（又称位置函数），即下式：

$$d = [-\ln(1-2r)]/2 \text{ 或 } r = (1-e^{-2d})/2 \qquad \text{（公式 8-7）}$$

8.5.3　TDT 分析与 ASP 分析

TDT 分析是 1993 年建立的一种基于系谱分析基础上的关联分析方法。主要是针对病例对照资料的关联分析（见下节）中因群体结构不同引起的误差进行纠正的一种统计方法。采用的资料中，必须具有双亲的样品，子女中至少有一名患者。通过对某一对等位基因的传递规律进行分析，分别对亲代中传递到患病子女的某种等位基因（A）和不传递到患病子女的某种等位基因（N）进行计数，然后利用下式计算 χ^2 值，自由度为 1，得到统计概率。此法又称 McNemar 检验。即：

$$\chi^2_{TDT} = \frac{(A-N)^2}{A+N} \qquad \text{（公式 8-8）}$$



如图 8-5，该家系中子女患病，双亲表型正常，同为杂合子。基因型 AA 为病例组，NN 为对照组，即需要比较 A 传递但 N 不传递组与 N 传递但 A 不传递组的 χ^2 值，根据式 8-8，图 8-5 家系 TDT 的 χ^2 值等于 $(2-0)^2/(2+0)=2$。如果分析样品中出现亲代资料缺失现象，需要用修正 TDT 法进行检验。

		不传递的等位基因	
		A	N
传递的等位基因	A	0	2
	N	0	0

图 8-5　TDT 分析示意图

当双亲资料完全缺失时，则须采用 ASP 分析，比较受累同胞与非受累同胞之间等位基因的分布频率。这是非参数连锁分析中常用的分析方法。ASP 分析中涉及两个基本概念，即血缘一致性（identity by descent，IBD）和状态一致性（identity by state，IBS）。IBD 是指两个同胞间相同的等位基因来源于同一亲代；IBS 是指

同胞间相同的两个等位基因分别来自父方和母方。有关 IBD 和 IBS 的示意图见图 8－6。家系 1 中 IBD＝1 但 IBS＝0；家系 2 中 IBD＝0 但 IBS＝1。通过分析子女中 IBD 和 IBS 的概率，可得知基因与疾病的相关性。在受累同胞对设计中，每个位点等位基因的 IBD 分为 0，1，2 三种情况，在不存在连锁的情况下，它们所占的比例分别为 0.25，0.5，0.25。如果某个位点与疾病相关，则该位点的 IBD 值更高。通过比较 IBD 等于 0，1，2 时的观察值（O）和期望值（E）的差别，计算 χ^2 值，自由度 df＝IBD 类型数－该基因位点的等位基因数。

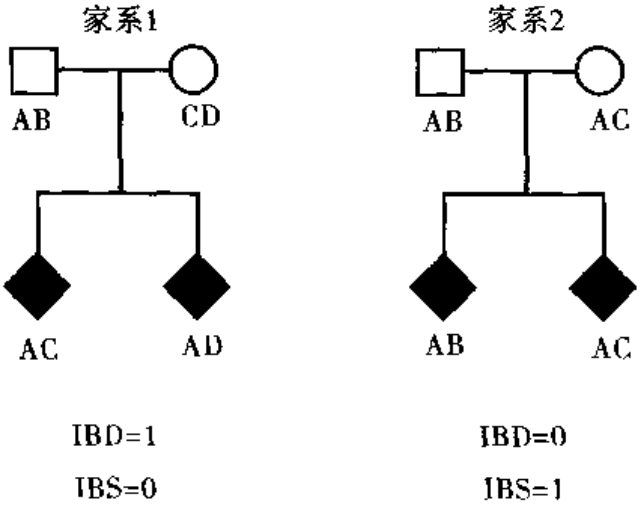


图 8－6　IBS 和 IBD 示意图

$$\chi^2_{IBD} = \Sigma (E - O)^2 / E \qquad \text{（公式 8－9）}$$

实际应用中，常用 MAPMAKER/SIBS 软件进行分析，也可用 GENEHUNTER 软件计算。值得一提的是，上述有关 LOD 分析、TDT 和 ASP 分析中都涉及 GENEHUNTER 软件的应用，该软件还可用于系谱分析、QTL 定位分析、单体型估计等，是目前连锁分析中最常用的软件之一。该软件由 MIT 的 Whitehead 基因组中心的 Eric Lander 小组由 1996 年设计出来，目前已更新到 3.0 版本，可通过 ftp 方法从该基因组中心的软件站下载，目前仅有供 Unix 操作系统的版本，尚不能在 Windows 支持下直接运行，也没有提供直接的服务器。用户必须在该中心注册登记。其操作手册可在 Rockefeller 大学的网站中得到（http://linkage.rockefeller.edu/soft/gh2）。

8.5.4　病例对照资料分析

基于群体的关联分析时，最常采用病例对照研究资料。这种关联性包括：①比较病例与对照间，易感候选基因或分子标记的等位基因频率的相对危险度或几率比值；②对于数量性状，如身高、血压等，比较病例与对照间该数量性状与患病率或发病率相关性的差别。

上述分析所涉及的统计方法参见 10.4 和 10.5 节。在判别关联分析统计结果时，需要对样品资料进行严格审核，无偏倚现象发生，同时所发现的关联基因或多态位点在功能上或染色体定位上具有生物学意义。如果分析结果具有统计学意义，在排除混杂因素（confoun-

der)和相互作用后,通常认为该等位基因与疾病的发生存在关联性,但也可能由于连锁不平衡现象的存在,该等位基因并非引起发病的直接因素。有关关联分析方法的实验设计还将在9.2.3节中讨论。

8.5.5 QTL定位[16~19]

前面所述的连锁分析和关联分析中,主要讨论的疾病属于二值性状(dichotomous trait),即有或无的两种相对情况。这通常是针对单基因病性状而言。多基因性状或数量性状的分析中,可采用回交(backcross)或杂交(intercross)的实验设计方法。然后对数量选择位点(quantitative trait locus, QTL)进行定位,简称QTL定位。所谓QTL,是利用统计方法对复杂性状进行分析的结果。QTL通常累及多个基因,也受环境因素影响。目前有关QTL定位方法中,常用的有单分子标记定位(single - marker mapping)、多分子标记定位(multiple - marker mapping)、区间定位(interval mapping, IM)、复合区间定位(composite interval mapping, CIM)和多区间定位(multiple interval mapping, MIM)等。这些方法的基本原理与前面所述的连锁分析相同,但需要进行计量资料的统计分析或多元相关与回归分析。

单分子标记定位分析,又称单分子标记回归分析,可采用正态分布的Z检验、t检验或F检验。这种方法相对简单,按该分子标记的不同基因型分组处理,统计分析不同组间表现型的差异。对于不同的分子标记,须逐个计算,反映差别是否具有统计学意义,但不能反映连锁关系的大小,即无法在全基因组水平进行扫描定位。如果考虑多个不同分子标记(或基因)可能都影响某一性状,则须建立多元回归模型进行分析,同时尽量考虑混杂因素和相互作用的影响。相关统计方法参见第10章。

区间定位方法是由Lander和Botstein于1989年建立的方法,其目的是将QTL定位在两侧翼分子标记(flanking marker)之间,同时增加检验效率。所用统计方法是似然性比值法(LRT),并通常计算Lods值。IM是目前QTL定位中最常用的方法,可以反映QTL与分子标记间的相对距离和连锁强度。复合区间定位是将IM与分子标记定位相结合的方法,其目的是考虑区间外其他分子标记对QTL的贡献。统计处理上是将LRT与多元回归结合。CIM可有效减低IM分析中区间内出现的双峰或多峰现象(又称"鬼峰"),但区间外若存在许多分子标记的影响,则减低检验模型中区间QTL的检验效率。

MIM是针对多个QTL的分析方法,与CIM不同的是,MIM同时进行多个区间的多元QTL分析,先使用MLE - EM进行参数的估计,再进行LRT分析。MIM可有效区分连锁QTL和相互作用QTL。这种方法计算过程较烦琐。

QTL定位中阈值(threshold value)的确定很重要,这与所选用的统计方法和样本量有关。通常,用不同方法检验时可得到相同的统计结论。目前QTL定位中常用的统计软件有SAS, MINITAB, R等软件,另外,也常使用一些遗传统计学软件如MapManager QTX, MAPMAKER/QTL, MAPMAKER/EXP, QTL CARTOGRAPHER等,须结合样本资料类型和QTL定位方法进行选用。

8.6 群体遗传学分析方法[11,12,20]

8.6.1 Hardy - Weinberg 平衡

在一个理想群体中(群体足够大、随机婚配、无自然选择、无新突变发生、没有群体的大规模迁移等),将等位基因分为野生型 N(正常)或突变型 A(致病)两种类型,三种基因型 NN,NA,AA 在群体中的分布保持平衡。英国数学家 Hardy 和德国遗传学家 Weinberg 得出了关于各基因频率和基因型频率的分布规律。等位基因 N(显性基因)的频率记为 p,等位基因 A(隐性基因)的频率记为 q,则基因型 NN,NA,AA 的频率分别为 $p^2,2pq,q^2$。则有:

$$p+q=1,(p+q)^2=p^2+2pq+q^2=1 \qquad \text{(公式 8-10)}$$

这就是著名的 Hardy - Weinberg 公式,又称遗传平衡定律,简称 H - W 定律。换句话说,在达到 Hardy - Weinberg 平衡的群体中,等位基因频率和基因型频率不再随世代发生改变。Hardy - Weinberg 平衡也适用于多个等位基因,即复等位基因。如果一个基因坐位上有 3 个等位基因,其平衡公式为:$(p+q+r)^2=p^2+q^2+r^2+2pq+2pr+2qr=1$。一般地,如果群体中某基因坐位上有复等位基因 $A_1,A_2,\cdots,A_m$,其对应的基因频率为 $p_1,p_2,\cdots,p_i,(i=1,2,\cdots,m)$,$f(A_iA_j)$ 表示各基因型频率($i<j$, $j=2,3,\cdots,m$),这样,等位基因 A_m 的频率为:

$$p_i=f(A_iA_j)+(1/2)\times\sum_{j=1}^{m}f(A_iA_j)\quad(i\neq j) \qquad \text{(公式 8-11)}$$

基因频率与基因型频率的关系为:

$$\sum_{i=1}^{m}(p_i)=\left[\sum_{i=1}^{m}(p_i)\right]^2=\sum_{i=1}^{m}p_i^2+2\times\sum_{i<j}^{m}(p_ip_j)=1 \qquad \text{(公式 8-12)}$$

公式 8-12 中,Σp_i^2 称为基因纯合度,$2\times\Sigma(p_ip_j)$ 称为基因杂合度,两者之和等于 1。Hardy - Weinberg 平衡公式常用于基因频率估计及杂合度的计算。在遗传流行病学中,常用 χ^2 检验公式判断该群体是否达到 Hardy - Weinberg 平衡,即 H - W 吻合度测验(goodness - of - fit test)。其计算公式为:

$$\chi^2_{H-W}=\Sigma(E-O)^2/E \qquad \text{(公式 8-13)}$$

式中,E 代表基因型的期望值;O 代表相应的观察值。自由度 df = 基因型类型数 - 该基因位点的等位基因数。

每一个体都是二倍体,基因型分为纯合子或杂合子。在实际应用过程中,Hardy - Wein-

berg 公式中的 p^2 表示显性等位基因纯合子频率，$2pq$ 表示杂合子频率，q^2 则为相应的隐性等位基因纯合子频率，对于常染色体隐性遗传病而言，q^2 也就代表疾病在群体中的发病率。通过查阅某疾病在人群中的发病率，可求得 q 值，再由 $p=1-q$ 算出 p，最后可求出 $2pq$，即携带者频率。例如已知某一群体中 PKU 的发病率是 1/10 000，且 PKU 是常染色体隐性遗传疾病。因此 $q^2=1/10\ 000$，$q=1/100$，所以，$p=1-q=99/100$，$2pq=2\times(1/100)\times(99/100)=0.019\ 8$。就是说该群体中 PKU 杂合子（携带者）的频率约为 1/50。

一般地，人类遗传病的发病率都很低，故 q 的数值非常小，p 约等于 1。因此，在运用 Hardy - Weinberg 公式进行计算时，可以用 $2q$ 来计算遗传病的基因携带者频率。如上面例子中 $2q=0.02$，与原来的 $2pq=0.0198$ 十分接近。

对于 X 连锁遗传病而言，男性只有一条 X 染色体，因此其基因型频率等于基因频率。即通过调查男性发病率，就是致病基因频率（q），或者通过调查女性患者（纯合子）的频率（q^2），其平方根就是致病基因频率（q）。例如红绿色盲是一种 X 连锁隐性遗传病，在某群体中男性发病率为 10%。则 $q=0.1$，$p=1-q=0.9$，女性患者频率为 $q^2=0.01=1\%$。女性杂合子（携带者）频率为 $2pq=2\times0.9\times0.1=0.18=18\%$。一般地，群体中某一 X 连锁隐性基因频率越小，男女发病率差异越大。在计算子女基因频率或基因型频率时，男性和女性应分别计算。

Hardy - Weinberg 公式不适于常染色体显性遗传病，这是因为杂合子都是患者且占患者的大多数，而纯合子患者非常罕见，甚至可以忽略不计。因此，所有的非患病者都是正常纯合子，而几乎所有的患者都是杂合子。

通常对于一个群体而言，突变和选择作用对遗传平衡的影响较小，在满足随机婚配的条件下，基本接近理想群体。移居可能对某些国家的局部地区显得重要，但通常不会带来明显影响。如果在选择配偶的过程中，发生近亲结婚，或由于有相同的地理因素、生理特点（如智力、身高、皮肤颜色等）以及特殊种族背景等原因而选择配偶，在这种情况下，Hardy - Weinberg 公式的应用难以得出正确的等位基因和杂合子频率，特别是在近亲结婚的情况下，偏差尤其明显。

8.6.2 近婚系数

近亲（或称亲缘关系）是指 3 ~ 4 代以内有共同的祖先。如果他们之间通婚，就称为近亲婚配。近亲婚配的夫妇有可能从他们共同的祖先那里获得同一基因，并将之传递给子女。如果这一基因按常染色体隐性遗传方式，其子女就可能因为是突变纯合子而发病。因此，近亲婚配降低了群体中杂合子频率，而增加了纯合子频率。近亲婚配使子女中得到这样一对纯合子或相同基因的概率，称为近婚系数（inbreeding coefficient，F）。下面以表兄妹之间的近亲婚配为例来说明近婚系数的一般计算方法。

图 8 - 7 是表兄妹近亲结婚的系谱。S 的近婚系数计算方法如下。假设 S 的共同祖先 P1 和 P2 某一常染色体基因坐位的等位基因分别为 A1、A2 和 A3、A4，即该位点共有 4 个等位基因。S 的近婚系数就是指表兄妹 C1 和 C2 婚配后，S 得到 A1A1，A2A2，A3A3，A4A4 的总概率。从图中可以看出，P1 将基因 A1 经 B1 和 C1 传给 S 需要经过 3 步，同样 P1 将基因 A1 经 B2 和 C2 传给 S 也需要经过 3 步，就是说，P1 的等位基因 A1 需要经过 6 步传递才能

使 S 的基因型成为 A1A1。每步传递的概率都是 1/2，因此，按照乘法原则，S 形成纯合子 A1A1 的概率为 $(1/2)^6$，同理，S 获得纯合子 A2A2，A3A3，A4A4 的概率也分别为 $(1/2)^6$，按照加法原则，S 的近婚系数为 $4\times(1/2)^6=1/16$。

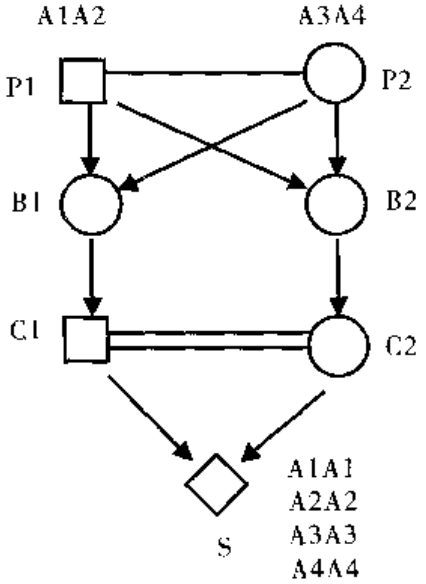


图 8-7　表兄妹结婚等位基因的传递和近婚系数计算

依此类推，可以将不同等级的近亲婚配中的近婚系数计算出来(表 8-3)。

表 8-3　近亲婚配类型及其近婚系数

近亲婚配类型	亲缘级别	近婚系数(F)
父母与子女之间	一级	1/4
同胞兄妹之间	一级	1/4
同父异母或同母异父的半同胞之间	二级	1/8
叔侄女之间	二级	1/8
舅甥女(或姑侄)之间	二级 1/8	
表兄妹之间	三级	1/16
堂兄妹之间	三级	1/16
半表兄妹之间	四级	1/32
二级表兄妹(从表兄妹)之间	五级	1/64

如果是 X 染色体上的基因，由于女性有两条 X 染色体，可能形成纯合子，而男性是半合子，不存在纯合的问题，近亲婚配对他没有影响。这样，在计算 X 连锁基因的近婚系数时，只计算女性的 F 值。从传递方式来看，男性 X 连锁基因一定传给女儿，概率为 1；而传给儿子的概率为 0，因此，X 连锁基因的近婚系数计算与常染色体基因有所不同。特别注意在传递通路上出现男到男时，该传递路径不予计算，即不考虑该基因纯合的可能性。

若为姨表兄妹，仍以图 8-7 谱系为例，作些适当修改。以 A1 基因代表 Y 染色体，A2 为 P1 的 X 染色体上等位基因，A3，A4 为 P2 的 X 染色体上两个不同的等位基因，S 为女性。

因此,S 的近婚系数实际上是指 S 获得 A2A2,A3A3,A4A4 纯合子的概率。A2 必然传给 B1 和 B2,经 B1 传给 C1 的概率为 1/2,C1 若获得 A2 必然传给 S,这样,A2 经 B1,C1 到 S 的传递只须计算 1 步,A2 经 B2,C2 传到 S 须计算 2 步,合计 3 步,即 S 获得 A2A2 的概率为 $(1/2)^3$。等位基因 A3 和 A4 的传递方式有所不同,经 B1,C1 的传递计为 2 步,经 B2,C2 的传递计为 3 步,合计 5 步,即 S 获得 A3A3 或 A4A4 的概率都是 $(1/2)^5$。S 的近婚系数为 $(1/2)^3+2\times(1/2)^5=3/16$。

近亲婚配对后代的影响,主要表现在增加隐性遗传病的发病率,而且先天畸形、早产和流产、幼儿夭折的风险也增加。在评价近亲婚配对群体的危害时,通常须在调查各类近亲结婚的基础上,计算出平均近婚系数(以 a 值表示)。a 值越大,对群体的危害越大。一般以 a 值为 0.01(即 1%)为高值。通常在发达、开放的社会中,a 值较低;在封闭、隔离或有特殊风俗的社会中 a 值较高。1980~1981 年,我国对北京地区的汉族人口调查表明,其近亲婚配率为 1.4%,平均近婚系数为 0.067%。

8.6.3 选择作用

绝大多数基因都受到自然选择(natural selection)的作用,具有某些等位基因或基因型的个体可能在子代中多保留下来,而另一些则少保留下来,即表现为群体中某些等位基因或基因型频率增加,另一些等位基因或基因型频率减少。但是若选择作用发生在育龄期之后,不会对后代产生影响。通常将某种基因型与其他基因型相比较时能存活并传递给子代的相对能力,称为适合度(fitness),用 f 表示。一般将正常纯合子个体的适合度定为 1,其他基因型的适合度根据相对生育率(relative fertility)来表示。例如在一项调查中,软骨发育不全(侏儒症的常见类型,由 FGFR3 基因突变所引起)患者 108 人,共生育后代 27 人,而患者同胞 457 人中,共生育 582 人,因此软骨发育不全的相对生育率为:$f=(27/108)\div(582/457)=0.196$。适合度为 0 时,表示患者通常不能婚育或后代无法存活。适合度还与环境的相互作用有关。

选择系数(selective coefficient)是用来表示在选择作用下适合度降低的程度,用 s 表示,即表示某种基因型在群体中不利于生存的程度,与适合度的关系为:$s=1-f$。选择作用对等位基因和基因型频率的影响因遗传方式不同而有差别。

对于常染色体显性基因,假设显性基因 A 的频率为 p,隐性基因 a 的频率为 q,选择对显性基因的选择系数为 s,则每一代显性基因 A 频率的改变 ΔP 为:

$$\Delta P=\frac{-sp(1-p)^2}{1-sp(2-p)} \qquad (公式\ 8-14)$$

通常群体中显性有害基因的频率很低,经过选择作用后,在没有新突变产生的情况下,这种显性有害基因经过若干代后,较容易从群体中消失。若突变和选择达到平衡,经推算可得知此时的突变率约为群体中患者发病率的一半,即约等于 p。

对于常染色体隐性基因,经选择后基因 a 的频率改变 ΔQ 为:

$$\Delta Q = \frac{-sq^2(1-q)}{1-sq^2} \qquad (公式\ 8-15)$$

群体中 q 值通常较小，每代基因 a 频率的降低约为 $sq^2(1-q)$，因此隐性致病基因的频率降低速率往往较缓慢，在没有新突变产生的情况下，最终仍会在群体中消失。若新的隐性突变的产生与选择达到平衡，此时突变率约等于 sq^2，当 $s=1$ 时，即为致死突变，突变率则等于患者发病率 q^2。

对于 X 连锁基因，因其性状通常在男性中表现出来，2/3 的 X 连锁基因存在于杂合子女性中，表型正常，不受选择的作用。对于 X 连锁隐性基因而言，群体中致病基因频率即等于发病率 q。在选择作用下，每一代的男性中将有频率为 sq 的致病基因被淘汰。在平衡群体中需要相同数目的突变基因来平衡，此时的新发生突变率为：

$$\mu = (1/3)sp \qquad (公式\ 8-16)$$

一般而言，许多突变是有害的，但大多数是中性的，仅极少部分突变为有害的。自然选择是一种保存有利突变和消灭有害突变的过程。值得注意的是，选择是一个复杂的生物过程，选择优势常与基因与基因、基因与环境的相互作用有关。

8.6.4 基因突变

基因突变是影响群体遗传平衡的重要原因。假设 μ 为等位基因 A 突变为 a 的突变率，P_0 为 0 世代 A 的基因频率，P_t 为 t 世代基因 A 的频率，在不考虑其他改变群体基因频率的因素时，基因频率的变化规律为：

$$P_t = P_0 \times e^{-\mu} \qquad (公式\ 8-17)$$

根据公式 8-17 可知，在仅考虑突变的影响时，等位基因 A 的频率随世代数的增加而呈指数规律下降。

8.6.5 遗传漂变

对于较小群体而言，通常难以维持遗传平衡，会出现基因频率的随机波动现象，称之为遗传漂变(genetic drift)。这在相对隔离的小群体中常见，表现为某些基因型较固定，另一些基因型减少甚至丢失的现象。遗传漂变是群体中出现某些新突变固定下来的原因，由此产生所谓的祖先效应(founder effect)。

假设某自然隔离群体有效容量为 N，通常为该群体个体数量的 1/3。某基因位点的等位基因总数为 $2N$，分别以 A 和 a 表示，其基因频率分别为 p 和 q 或 $(1-p)$，经过一代后，基因频率改变的方差可表示为：

$$\sigma = pq/2N \qquad (公式\ 8-18)$$

对于一个有效容量为 N 的群体，若无选择作用，所产生的新突变从该群体中丢失的概率为$(1-1/2N)$。若新突变能固定下来，则形成祖先效应。群体越小，新突变被固定下来的概率越大。

8.7 遗传风险计算[11,21]

再发风险率是指曾生育过一个或几个遗传病患儿，再生育该病患儿的概率。现在这一概念已扩展到凡有信息可导致一对夫妇生育某种遗传病患儿(包括第一胎)的概率，又称患病风险或遗传风险。遗传咨询中，一般分单基因病再发风险率、多基因病再发风险率、染色体病再发风险率、线粒体病的再发风险率等。

8.7.1 单基因病再发风险率的计算

单基因病的基因型已明确者，再发风险率按孟德尔定律计算(参见2.3.1节)。一般遵循以下原则。

8.7.1.1 常染色体显性遗传病

此类疾病患者一般为杂合子，纯合子多于胎儿期或婴幼儿期死亡，或于幼年期发病后很难存活到结婚年龄。因此夫妇一方为患者时，子代每胎的再发风险率为1/2。夫妇双方同为患者的机会极少见，这时子代再发风险率为3/4。

8.7.1.2 常染色体隐性遗传病

患者一般见于其双亲同为杂合子的情形，再发风险率为1/4；杂合子出现概率为1/2，完全正常的概率为1/4；或者说，表型正常的子代概率为2/3。

8.7.1.3 X连锁隐性遗传病

母亲为携带者，与正常男性婚育子女中，男孩患病风险为1/2；女孩有1/2为携带者，1/2为完全正常，即所生女孩表型全部正常。若父亲为患者，与正常女性婚育子女中，男孩全部正常，女孩都是杂合子。母亲为患者，与正常男性婚育子女中，男孩全部患病，女孩全部为杂合子。

孟德尔遗传病中，如果夫妇双方或一方的基因型不明，再发风险率的计算必须依靠Bayes分析。例如一位女性，其母亲是甲型血友病携带者，父亲正常，她本人表型正常，是该病携带者的概率为1/2。目前她已生育了2个正常男孩，表明是携带者的可能性减小(即$<1/2$)，确切的概率是多少？这就需要利用Thomas Bayes提出的著名的逆概率定理来计算，通常称为Bayes分析。主要包括以下基本概念。

(1)根据孟德尔遗传规律所得到的初始概率，称为前概率(prior probability)，记作$P(A)$，不发生A的概率记作$P(\bar{A})$。

(2)如果事件B的概率$P(B)$与早先出现过的事件A有关联，则在发生事件A的条件下，再发生事件B的概率称为条件概率(conditional probability)，记作$P(B|A)$。计算公式为：$P(B|A)=P(AB)/P(A)$。

(3)上式中 $P(AB)$ 表示事件 A 和 B 同时发生的概率,称为联合概率(joint probability),即等于前概率与条件概率的乘积。

(4)在考虑上述几类概率基础上所得到的最后概率,称为后概率(posterior probability)。即计算在 B 发生的条件下,A 发生的概率 $P(A|B)$ 和 A 不发生的概率 $P(\bar{A}|B)$。

后概率的计算是通过下面的 Bayes 公式:

$$P(A|B)=\frac{P(A)\times P(B|A)}{P(A)\times P(B|A)+P(\bar{A})\times P(B|\bar{A})} \qquad (公式\ 8-19)$$

$$P(\bar{A}|B)=\frac{P(\bar{A})\times P(B|\bar{A})}{P(A)\times P(B|A)+P(\bar{A})\times P(B|\bar{A})} \qquad (公式\ 8-20)$$

在遗传病再发风险的计算中,Bayes 分析的目的是将所掌握的致病基因 A 的孟德尔遗传方式,根据家系中有关成员间的关系、实验检测方法的灵敏度和特异性等所有资料(条件 B)综合起来,计算出这种特定 B 条件发生后,某个个体携带致病基因 A 的概率,即 $P(A|B)$,从而得出遗传病发生的风险率。实际应用中,为方便起见,常通过列表的方法比较直观地推导出后概率(表 8-4)。

表 8-4 Bayes 分析表

	是 A 基因携带者	不是 A 基因携带者
前概率	$P(A)$	$P(\bar{A})$
条件概率	$P(B|A)$	$P(B|\bar{A})$
联合概率	$P(A)\times P(B|A)$	$P(A)\times P(B|\bar{A})$
后概率	$P(A|B)$	$P(\bar{A}|B)$

回到上面所举的例子中,现利用 Bayes 分析表(表 8-5)来计算该女性生育两胎正常男性后,是甲型血友病携带者的概率。按孟德尔遗传规律,该女性是携带者和不是携带者的概率均为 1/2(前概率)。当她是携带者时,两个儿子都正常的条件概率为 1/2 × 1/2 = 1/4,此时联合概率为 1/2 × 1/4 = 1/8,依照表 8-5 的 Bayes 分析表可以得出她为携带者的后概率为 1/5,较前概率的 1/2 明显下降。理论上,她每多生下一个正常男胎,其为携带者的概率相应地减小。然而,一旦生育了一位患血友病的男孩,那么她就成为肯定携带者,从而推翻了以往所得出的任何后概率计算结果。

表 8-5 X-连锁隐性遗传 Bayes 分析表

	Ⅱ2 是血友病基因携带者	Ⅱ2 不是血友病基因携带者
前概率	1/2	1/2
条件概率	1/2 × 1/2 = 1/4	1
联合概率	1/2 × 1/4 = 1/8	1/2
后概率	(1/8)/(1/8 + 1/2) = 1/5	(1/2)/(1/8 + 1/2) = 4/5

8.7.2 多基因病再发风险率的计算

多基因病在家系中的再发风险与多种因素有关(参见2.3.2节),其再发风险率的估计一般根据以下几方面综合评估。

8.7.2.1 群体发病率与遗传度

多数多基因病的群体发病率为0.1%~1%,遗传度介于70%~80%。这种情况下可用Edward公式来估计发病风险,即$f=\sqrt{P}$。其中f为患者一级亲属发病率,P为该群体发病率。例如我国人群中,唇腭裂的发病率为0.17%,遗传度为76%,因此,患者一级亲属的发病率为$\sqrt{0.001\ 7}$,约等于4%。如果群体发病率高于1%或遗传度高于80%,则患者一级亲属发病率要高于$\sqrt{P}$;反之,如果群体发病率低于0.1%或遗传度低于70%,则患者一级亲属发病率要低于$\sqrt{P}$。

8.7.2.2 一级亲属患病人数及亲缘关系级别

一个家庭中患病人数越多,发病风险越高。一般地,如果有两位一级亲属患病,经验风险率就相当于原来的1倍;如果有三位一级亲属患病,再发风险约为原来的3倍。例如,一对夫妇生育一个唇腭裂患者后,再发经验风险为4%;如果再出生一个患儿,则下次再发风险增加2~3倍,接近10%。另外,随着亲缘关系级别降低,发病风险也相应降低。

8.7.2.3 疾病的严重程度

病情越严重,其易患性更趋近阈值,因而再发风险更大。例如双侧唇/腭裂的再发风险较单纯性单侧唇裂为高。散发性双侧唇裂的经验再发风险率为8%,而散发性单侧唇裂的再发风险率只有4%。这是因为病情越严重,表明携带更多的易感基因,其平均易患性就越高,所以生育患病后代的可能性也越大。

8.7.2.4 患病者性别

当一种多基因病的群体发病率表现出性别差异时,表明不同性别的易患性不同。群体发病率高的性别,其子女发病风险低;群体发病率低的性别,其子女的发病风险高,称之为Carter效应。因为在群体发病率低的性别,一旦发病,表明其所具有的易感基因更多,才有可能超过阈值而发病,因此其子女的发病风险更高。

8.7.2.5 近亲婚配

通常多基因病近亲婚配的后代发病风险略高,但不如单基因病明显。

8.7.2.6 经验风险率

由于上述这些因素之间的关系复杂,所以对多基因遗传病的再发风险的评估通常以经验风险率(empiric risk)表示,并可以通过查表得到,也可通过专用电脑软件计算后得到再发风险率。

8.7.3 染色体病再发风险率的计算

在评估染色体病再发风险时一般遵循以下几点。

(1)如果父母的核型正常,子女中有染色体数目异常的染色体病患者,再生育子女的再发风险与正常群体发病率相似。

(2)双亲之一为染色体平衡易位携带者,其子女再发风险明显升高。以 14/21 平衡易位型 21 三体携带者为例,其后代活产子女中理论上有 1/3 为正常,1/3 为易位型 21 三体综合征患者,1/3 为平衡易位型携带者。

(3)双亲之一为染色体病患者,由于多数染色体病患者不能生育,因此子代再发风险率较难预测,一般都高于群体发病率。如 21 三体综合征女性患者如能生育,其子代再发风险率约为 1/3。

(4)随着母亲妊娠年龄增大,生育染色体病(尤其 21 三体综合征)的风险逐渐升高。

(5)当一位孕妇有 21 三体(或 13 三体、18 三体)妊娠史阳性时,其再发风险会升高。这是因为:①这些妇女生殖细胞减数分裂过程中的染色体不分离,易感性高;②夫妇一方可能是低水平性镶嵌体携带者。通常,凡是有过 21 三体、13 三体或 18 三体妊娠史的女性,其再次生育患同样异常患儿的风险会比同年龄对照组的高 0.5%。如果将 35 岁以下孕妇归为一组,而将 35 岁和 35 岁以上的归为另一组,那么属前一组而且有 21 三体综合征妊娠史的孕妇再次生育 21 三体胎儿的风险为 1%,而属后一组者则应根据其年龄查阅有关风险率计算表得出结果。

(6)不平衡性染色体病再发风险的估计:患不平衡性染色体病的婴儿给家庭和社会带来的影响比因流产而失去胎儿的影响更为严重。这是因为当染色体片段处于不平衡状态时,基因组也随之失去平衡而导致胚胎或胎儿生长发育的障碍。故单体或三体异常都容易造成机体畸形的发生。并非所有的染色体畸变都会导致不平衡性染色体病患儿的出生。通常,部分性单体带来的危害比部分性三体的要大。在评估生育不平衡性染色体病子女的风险时,一般要从三个方面进行考虑:①单体片段越大,胎儿宫内存活机会越小;② 在单体片段和三体片段同时存在时,单体片段的效应通常占优势;③不平衡片段越小,胎儿存活至分娩或存活出生的机会越大。

8.8 亲子关系排除与肯定概率[22]

通过对人类基因组的解析,发现人与人之间 99.99% 的碱基序列都相同,就是说两个不同个体间平均每一万个碱基中大约有一个碱基的差别。正是这些 DNA 序列上细微的差别,使得每个人都具备区别于他人的碱基序列标签,又称 DNA 多态性,也是个体差异形成的根本原因。同时根据遗传学传递规律,子女的身份标记必定来自于双亲,据此可用于亲子关系鉴定。目前最常用做个人身份的 DNA 标签是 DNA 短串联重复序列(STR)。以现行国际上常用的 9 个 STR 位点为例(美国 ABI 公司),应用遗传分析仪进行基因型分析,个人识别率达到 1.05×10^{-10}。即理论上,通过检测这 9 个 STR 位点,可将 1 个人从约 100 亿人中鉴别出来,亲权关系排除率可达 99.98% 以上。这一有力的个体识别和亲权鉴定技术手段,为司法审判、侦察破案提供了科学依据。DNA 身份标记是目前亲子鉴定中最准确的一种,较以往的血型分析和 HLA 分型鉴定方法可靠性高出 10 到 100 倍。如果小孩和被测男子的 DNA 模式在一个或多个 STR 位点不吻合,被测试男子便被完全排除他是亲生父亲的可能性。实际应用中,对于亲子关系分析,通常需要给出相应的肯定和排除概率,得出科学而客观的判断。

亲子关系累积肯定概率的计算公式为：

$$W = PI/(PI+1),\ PI = \prod_{i=1}^{n}(PI_i) \qquad (公式8-21)$$

式中，PI_i表示亲权指数(paternity index)，等于携带生父基因的概率(X)/生父基因的频率(Y)。实际应用中，还可根据条件概率，按 Bayes 公式计算 W 的后概率。一般认为，当 W > 99.73%，可完全肯定亲子关系的存在；当 W <0.27 时，可完全排除亲子关系，但须同时考虑排除概率的大小。亲子关系累积排除概率的计算公式为：

$$P = 1 - \prod_{i=1}^{n}(1-p_i) \qquad (公式8-22)$$

式中 p_i表示各多态位点的排除概率，与等位基因数目和亲代数目有关，详见参考文献[22]。

(李 巍)

参 考 文 献

1 FreeMan S & Herron JC. Evolutionary Analysis. 2nd ed. New Jersey: Prentice Hall, 2001. 437 ~464

2 Kimura M. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. J Mol Evol, 1980, 16:111

3 Hershkovitz MA, Lewis LA. Deep - level diagnostic value of the rDNA - ITS region. Mol Biol Evol, 1996, 13:1 276

4 Kumar S, Tamura K, Nei M. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci, 1994, 10:189

5 Felsenstein J. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol, 1997, 46:101

6 Rzhetsky A, Nei M. Statistical properties of the ordinary least - squares, generalized least - squares, and minimum - evolution methods of phylogenetic inference. J Mol Evol, 1992, 35: 367

7 Huelsenbeck JP. The robustness of two phylogenetic methods: four - taxon simulations reveal a slight superiority of maximum likelihood over neighboring joining. Mol Biol Evol, 1995, 12: 843

8 Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol, 1981, 17: 368

9 Jow H, Hudelot C, Rattray M, Higgs PG. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol Biol Evol, 2002, 19:1 591

10 Huelsenbeck JP, Larget B, Miller RE, Ronquist F. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol, 2002, 51: 673

11 李巍, 何蕴韶. 遗传咨询. 郑州: 郑州大学出版社, 2003. 57~76

12 Rimoin DL, Connor JM, Pyeritz RE, *et al*. Emery and Rimoin's Principles and Practice of Medical Genetics. 4th ed. London: Churchill Livingstone, 2002. 410~438

13 Daly AK. Candidate gene case-control studies. Pharmacogenomics, 2003, 4: 127

14 Schulze TG, McMahon FJ. Genetic association mapping at the crossroads: Which test and why? Overview and practical guidelines. Am J Med Genet, 2002, 114: 1

15 Clark AG. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev, 2003, 13: 296

16 Yang MCK. Introduction to Statistical Methods in Modern Genetics. Amsterdam: Gordon and Breach Science Publishers, 2000. 43~194

17 Broman KW. Review of statistical methods for QTL mapping in experimental crosses. Lab Animal, 2001, 30: 44

18 Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet, 2001, 3: 43

19 Sen S, Churchill GA. A statistical framework for quantitative trait mapping. Genetics, 2001, 159: 371

20 陈竺. 医学遗传学. 北京: 人民卫生出版社, 2001. 236~253

21 Young L. Introduction to risk calculation in genetic counseling. 2nd ed. New York: Oxford University Press, 1999. 10~14

22 江三多, 吕宝忠. 医学遗传数理统计方法. 北京: 科学出版社, 1998. 339~363

第9章　药物基因组学与化学信息学

药物遗传学(pharmacogenetics)和药物基因组学(pharmacogenomics)是两个易于混淆的概念。通常认为后者是从前者演化而来。药物遗传学已有40多年的历史,而药物基因组学是基因组学的一个分支(参见第3章),是近10年来随基因组学的发展而兴起的。药物遗传学主要研究药物与个体特征的相互作用,通过对各种表型的研究,如药物在个体中的有效性、毒性、耐受性等,探讨这种个体差异的产生与个体特征性生物标志物(基因多态性)的关联性,从而预测药物在个体的反应性。简而言之,药物遗传学就是研究决定药物反应的遗传性状的个体差异,即个体的基因组成特点如何决定药物的有效性和安全性。药物基因组学是在全基因组水平研究药物或毒物的效应,即系统研究药物如何影响机体的基因表达。与药物遗传学不同的是,药物基因组学重点针对不同药物对全套基因组表达的差别,从而确定药物的反应性,而不是针对同一种药物在不同个体中反应性的差别[1]。两者的对照参见表9-1。由此可见,药物基因组学将涉及更多、更复杂的生物信息学方法。尽管这两者的区别非常明显,但实际应用中,药物遗传学和药物基因组学的概念常相互交换使用,本章将涵盖这两部分的内容。

化学信息学(chemoinformatics)的概念于1998年首次提出,可看做在生物信息学和药物基因组学基础上发展的一个分支学科,也是化学和信息学基础上产生的交叉学科,包括药物的分子设计、分子靶向、分子模拟和相关数据库等。目前化学信息学与生物信息学并驾齐驱,成为许多制药企业、生物技术公司、临床化学研究中重要的工具。其主要应用于以下几方面:①根据用户要求设计化学分子结构,从相应的数据库中查找有关的结构信息;②利用图形显示,比较结构的相似或相同性;③储存和检索有关化学分子的二维或三维结构,并模拟分子间的相互作用。

表 9－1　药物遗传学和药物基因组学的比较

	药物遗传学	药物基因组学
研究策略	研究药物个体反应性的差别，通常受遗传性状控制	不同药物对基因表达谱的影响，通常在全基因组水平评估
研究方法	主要检测 SNP 为主	主要检测 mRNA 表达谱
研究目的	为患者提供更好的治疗方案	药物筛选和药物设计
研究模式	一种药物，多个基因组	多种药物，一个基因组
研究重点	个体差异性	药物差异性

9.1　药物个体反应性的遗传基础[1,2]

药物反应性(drug response)是指机体对药物反应的程度，主要包括药物的有效性和毒性(或称安全性)。通常通过监测血药浓度结合表型观察来评价药物的反应性。血药浓度分为治疗失效、治疗浓度、毒性水平等不同浓度，这由药物的药效学(pharmacokinetics)和药动学(pharmacodynamics)所决定。

本章第三节有关药物分子设计中将介绍其中有关内容。

药动学主要研究机体如何作用于药物，包括药物吸收、结合与分布、生物转化或代谢，以及排泄等过程。药效学主要反映药物如何作用于机体各组织器官，包括药物与受体的结合和信号传导过程。图 9－1 示意药效学和药动学的不同分期。药物代谢是药动学的重要环节，目前了解得比较清楚，分Ⅰ期(功能化)和Ⅱ期(结合)两个阶段，参见图 9－2。研究药动学和建立一些数学模型，这也被纳入广义的生物信息学范畴。目前已有相应软件问世，如 ADAPT 软件(version Ⅱ，1997)。

常见的药动学遗传差异表现在药物代谢酶类(drug－metabolizing enzyme，DME)，常见的药效学遗传差异则表现在药物受体、转运蛋白和离子通道等。这种差别在不同个体中可相差 10～40 倍，因而又称药物个体反应性。通常区分为反应型(responder)和非反应型(non－responder)，或称敏感性(drug－sensitive)和耐受性(drug－resistant)。这往往涉及多个不同基因，同时也受药物之间相互作用以及环境因素的作用等影响。

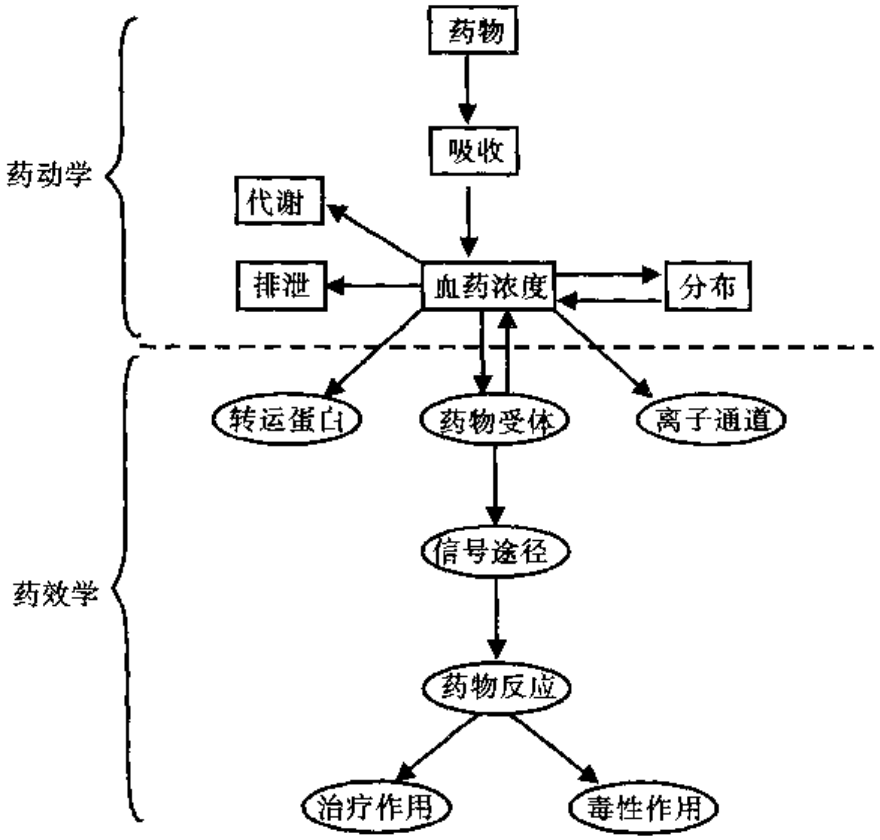


图9-1　药动学与药效学过程示意图

9.1.1　药物遗传学疾病概述[1,2]

基因组多样性(参见3.1.4节)是药物反应性的遗传基础。就是说,由于每个个体的遗传组成(make-up)或基因组不同,表现为对同一种药物的药动学和药效学上的差异,因而表现出对药物反应性的不同。对某些个体而言的治疗剂量可能是另一些个体的中毒剂量,或表现出副作用(adverse drug reaction,ADR)。根据美国1966~1996年的统计资料,严重和致死性的药物副作用分别占6.7%和0.32%,占死因的第四到第六位,超过因肺炎和糖尿病引起的死亡人数。因ADR住院患者占总住院人数的7%。副作用的表现形式有多种,可以是剂量依赖性,也可以是剂量不依赖性(如过敏反应),还可以是剂量与时间双重依赖性(累积效应)、给药时间性、药物撤退性或无法预见的治疗失败。副作用的大小依个体不同而有差异,主要受遗传因素控制。临床遗传学将那些与药物遗传性状相关的疾病归类为药物遗传学疾病。表9-2所列为常见药物遗传学疾病或性状。另外,其他一些非遗传因素也是药物副作用差异的原因,如肾、肺、肝、心功能状况,年龄与性别,药物之间的相互作用,环境和营养因素,疾病发病和严重程度等。

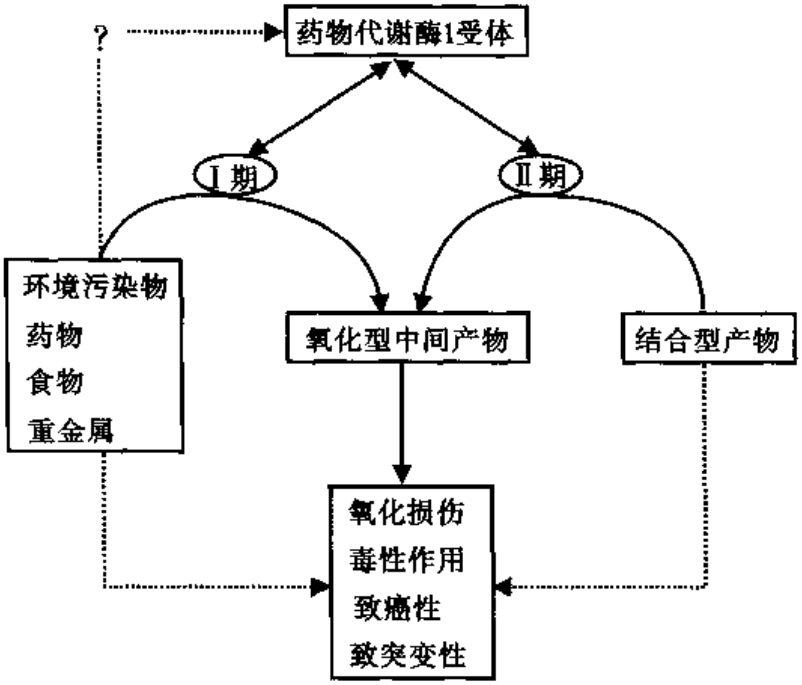


图 9－2　药物代谢酶类/受体作用的分期

表 9－2　常见药物遗传学疾病

A. 酶活性或蛋白改变	B. 受体、转运蛋白或离子通道的改变
N－乙酰转移酶(NAT2，NAT1)	苯硫脲尝味反应
葡萄糖－6－磷酸脱氢酶(G6PD)缺乏症	香豆素抗凝功能抵抗
高铁还原血红蛋白血症与血红蛋白病	钾离子通道(KVLQT1，KCNMB2 等)与长 QT 综合征
NADH－细胞色素 b_5 还原酶	雄激素不敏感症
P450 单胺氧化酶(CYP)	雌激素不敏感症
谷胱甘肽转移酶(GST)	多种特异性有机阴离子转运蛋白(MOAT)与 Dubin－Johnson 综合征
磺酸转移酶(SULT)	胰岛素耐受症
硫代嘌呤甲基转移酶(TPMT)	库欣综合征
硫甲基转移酶(THMT)	视紫质变异与色素性视网膜炎
儿茶酚 O－甲基转移酶(COMT)	加压素耐受症(AVPR2)
组胺 N－甲基转移酶(HNMT)	钙离子通道纳洛定受体(RYR1)与恶性高热
二乙基对硝基苯磷酸酯酶(PON1)	维生素 B_{12} 内在因子缺乏症
UDP 葡萄糖醛酸转移酶(UGT1A1)与 Gilbert 氏病	β－肾上腺素能受体(ADRB1，ADRB2，ADRB3)与哮喘病的β－受体拮抗剂治疗
NAD(P)H 苯醌氧化还原酶(NQO1)	血管紧张素Ⅱ受体与 ACE 抑制剂反应性
微粒体可溶性环氧化物水解酶(EPHX1，EPHX2)	磺胺脲受体(ABCC8)与磺胺脲反应性
非典型乙醇脱氢酶(ADH)与酒精中毒	5－羟色胺受体(HTR)及其转运蛋白(SLC6A4)与氯氮平的反应性
乙醛脱氢酶(ALDH2)与酒精中毒	多巴胺受体(D2DR，D3DR，D4DR)及多巴胺转运蛋白(SLCG3)与精神类药物反应
α_1－抗胰蛋白酶(AAT)	β－chemokine 受体(CCR5)与 HIV 感染的抵抗性
α_1－抗糜蛋白酶(ACT)	胆固醇转移蛋白(CETP)与降胆固醇药的作用
血管紧张素转换酶(ACE)与 ACE 抑制剂中毒	多种耐药转运蛋门(MDR1，MDR3)与化疗药物耐受性
过氧化氢酶与低过氧化氢酶血症	氨基酸转运蛋白(SLC3A1)与胱氨酸尿症
二氢嘧啶脱氢酶(DPYD)与氟尿嘧啶毒性反应	
丁酰胆碱酯酶(BCHE)与琥珀酰胆碱所致呼吸暂停	
P53 蛋白与 5－氟尿嘧啶诱导的细胞凋亡	
表面活性蛋白缺乏与肺的毒性	
乳糖不耐受症	
二氢叶酸还原酶过度表达与 5－氟尿嘧啶的耐药性	
c－ERBB2 过度表达与乳腺癌的早期治疗	

9.1.2　药物代谢酶类[1]

主要指肝脏中参与药物Ⅰ期和Ⅱ期作用的酶类。这些代谢酶相关基因的突变可导致酶活性丧失，因而表现为代谢底物的堆积，这些个体被称为“弱代谢者”（poor metabolizer），易于出现药物底物的堆积，产生毒性反应。由于基因重复，代谢酶活性增强，使得某些药物的代谢过于敏感，这些个体被称为“强代谢者”（ultra - extensive metabolizer），常表现为对药物的耐受性，常规剂量下表现为治疗失败。一些位于基因编码区或调节区的SNP，也可使酶的结构或构象发生一定改变，表现出酶活性增强或减弱，出现对药物的反应性不同。目前药物遗传学研究的重要内容之一就是在这些代谢酶水平，探讨其基因型与表现型的关系。

参与药物Ⅰ期作用的酶专司功能基团的修饰，包括水解、氧化、还原、羟基化等。绝大部分是通过细胞色素P450同工酶（CYP）的作用。人体中，已发现约80种CYP，每种CYP的催化特异性有所不同，其调节作用也不同，其中仅5～6种CYP酶负责绝大部分药物的代谢。这些酶当中，CYP2D6，CYP3A4，CYP2C19的多态性高，是迄今为止ADR研究中最多的3种酶。CYP2D6参与30多种药物的代谢。已知至少16种基因变异导致无义突变、氨基酸替换、微卫星重复数目改变、基因重复或缺失等，导致酶活性的减弱或增强。例如中国人（约30%）中常出现的所谓金雀花碱（sparteine）－异喹胍（debrisoquin）等位基因，导致该酶活性丧失，治疗常规剂量下易于出现对氯丙嗪、氟哌啶醇、可待因、抗心律失常药物的副作用或中毒反应。CYP3A4参与大部分药物的代谢并影响其在小肠中的吸收。已报道多种基因突变可影响肝脏中该酶的活性，其中较常见的是启动子区的基因多态性。CYP3A4与药物反应性的关系尚未完全阐明。与CYP2D6一样，该基因的多态性在不同人群中有所不同。CYP2C19的多态性可影响对质子泵抑制剂（抗溃疡药）奥美拉唑（omeprazole）的代谢，常规剂量下约20%的欧洲高加索人和更高比例的亚洲人中，表现对该药的耐受性。

参与药物Ⅱ期作用的酶是将内源性的极性分子与药物相连，以利于药物的排泄。一些酶的基因变异可影响个体对某些药物的反应性。如尿嘧啶二磷酸葡萄糖醛酸转移酶1A1（UGT1A1），N－乙酰转移酶2（NAT2），硫代嘌呤S－甲基转移酶（TPMT）等。UGT1A1参与亲脂性化合物和内源性物质如胆红素的解毒。启动子区的多态性可改变该酶的表达量，表现为对一大类药物代谢的改变。NAT2基因的纯合性突变导致该酶活性的丧失，产生剂量依赖性的药物毒性反应，如肼苯哒嗪引起的红斑、异烟肼引起的神经损害、磺胺药物引起的过敏反应等。TPMT的多态性可导致硫代嘌呤类药物如咪唑硫嘌呤、6－巯基嘌呤的生物转化，在某些病人中可导致造血系统致命的毒性反应。

9.1.3　转运蛋白[1]

药物在体内的沉积除肝脏代谢和肾小球滤过作用外，也与药物的吸收和穿透组织的作用有关。吸收和转运过程涉及一系列膜转运蛋白（transporter），这些膜转运蛋白与溶质和体液平衡的维持有关。据估计，人类基因组中含500～1 200种转运蛋白，与药物转运有关的蛋白了解得最清楚的是多种耐药蛋白（multidrug resistant protein，MDR），已明确的其他药物转运蛋白有多种耐药相关蛋白（MRP）、有机阴离子转运多肽（OATP）、有机阴离子转运蛋白

(OAT)、有机阳离子转运蛋白(OCT)等,可查阅www.med.rug.nl/mdl/tab3.htm。

已发现的 MDR 有 MDR1,MDR3 和 SPGP。其中研究最多的是 MDR1,编码 P-糖蛋白,具有多态性。这种转运蛋白将许多结构不同的药物,例如许多肿瘤化疗药物、HIV 蛋白酶抑制剂、CYP3A4 代谢产物等,泵出细胞外,送回到肠管中。MDR1 的作用物通常是亲脂性或为两性分子,含有一个或多个芳香环。第 26 外显子的一种多态性(3435C>T)可导致小肠中 P-糖蛋白减少,与地高辛的慢吸收有关。同时具有 TT 基因型者较 CC 或 CT 基因型者血浆地高辛浓度较高,这是导致个体 ADR 的主要原因。

还有一类参与药物转运的蛋白又称血浆结合蛋白,主要包括血清白蛋白、α_1-酸性糖蛋白。这些蛋白的多态性也可导致药物反应的个体差异。如血清白蛋白的两种变异体,Canterbury (Lys313Asn)和 Parklands (Asp365His)可导致血清白蛋白与华法林(warfarin)、水杨酸盐、安定等药物的亲和力降低。α_1-酸性糖蛋白的基因变异可导致与华法林、丙咪嗪等药物的结合力下降,使血浆中游离的药物浓度增加。此外,一些离子通道蛋白如钾离子通道、钙离子通道等也参与药物在体内的转运与分布。

9.1.4 药物受体[1]

许多药物与特异性靶蛋白相互作用,表现出其药理作用。这些靶蛋白主要包括膜受体和某些酶蛋白,如 β_2-肾上腺素能受体、胰岛素受体、各种含血清素的受体、血管紧张素转换酶(ACE)、羟甲戊二酰辅酶 A 还原酶等。此外,某些药物还可作用于信号转导蛋白,如过氧物酶体增殖物激活受体 γ-2 (PPAR-γ-2)。这些药物靶蛋白常表现出多态性,改变对药物的敏感性,即药动学的改变。这些靶分子也常与疾病的易感性有关,如 PPAR-γ-2 与Ⅱ型糖尿病、胆固醇酯转移蛋白(CETP)与冠状动脉粥样硬化、β_2-肾上腺素能受体与哮喘、ACE 与原发性高血压等。

多数情况下,影响 ADR 的多态性与上述药物靶向结合位点多态性不同。例如,凝血酶原基因 3'非编码区的多态性可导致妇女服用口服避孕药后脑血栓形成的危险性增加。再如,由于 ranodine 受体基因的突变,使用麻醉剂如氟烷或琥珀酰胆碱后,可出现高热反应。也有部分药物的 ADR 多态性与药物靶向结合位点多态性相同,如多巴胺 D_3 受体的多态性与服用安定类药物引起的迟缓型运动障碍有关,这种多态位点也是药物作用的靶位点。

9.2 药物遗传性状的鉴定方法[1]

目前已得到公认的是,药物反应性或药物遗传学疾病是一种多因素复杂性状,又称药物遗传性状。因此,药物遗传性状的分析方法依照 8.5 节中多基因性状分析方法。两种基本的研究方法是候选基因研究(candidate gene studies)和全基因组连锁不平衡作图(whole genome linkage disequilibrium mapping)。这两种方法各有其优缺点。候选基因研究需要有可靠的疾病基因或药物反应相关基因的资料,才有可能找到这种关联性。全基因组 LD 作图往往需要样品量大,而且 SNP 基因分型方法准确可靠、成本低、高通量。实际应用中,这两种方法常联合使用,以相互取长补短,即通常先采用全基因组扫描方法,找到关键区,再根据

关键区内的基因，筛选候选基因，称之为定位候选克隆方法。这些研究中的关键是确定所需要的 SNP 标记物种类和数量，以及根据 SNP 的连锁不平衡性，确定 SNP 单体型的区间大小和人群中的频率，相关问题的讨论参见 3.5.2 节。由于基于 SNP 单体型的全基因组作图方法尚在建立之中，而且目前的应用成本较高，因此，现阶段主要依靠候选基因研究的方法来分析药物遗传性状。常用关联分析方法（association study）来研究 SNP 与药物遗传性状的关联性。基因与药物的相互作用需要进一步利用药物基因组学方法加以证实。

9.2.1 候选基因研究中的病例对照分析[1-3]

候选基因研究主要根据以往的实验资料或知识，探讨药物反应性与那些药动学、药效学、病理学相关的基因之间的关系。通常是利用病例对照资料，分析若干候选基因的 SNP（主要为 cSNP）与药物反应性的关联性（association）或相关性（correlation）。统计方法参见第 10 章。

病例对照资料的候选基因研究中，有以下几个值得注意的问题。

9.2.1.1 SNP 的基因型频率和 SNP 的数量

人群中当某个感兴趣的 SNP 基因型频率小于 0.2 时，对结果判断比较困难，除非所得到的几率比值特别高。对于这种罕见、少见基因型的分析，即使增加样品量，帮助也不大。解决的办法之一是，选择多个 SNP 进行分析，又称多种分析（multiple testing）。多种分析时常需要利用 Bonferroni 不等性（Bonferroni inequality）原理进行校正，校正公式为：

$$P* = 1 - (1 - p)^n \qquad \text{（公式 9-1）}$$

式中，$P*$ 表示总概念；p 表示观察概率值，即待校正概率值；n 表示检验假设的数目。当几个 SNP 之间存在 LD 时，该校正公式不适用，即可能出现假阴性。

9.2.1.2 病例和对照的选择

病例对照研究中的病例和对照的选择较关键，其设计方法下文将叙及。通常认为所选病例和对照之间，要求各组种族、性别、年龄等最好相匹配。如果出现明显的人群分层现象（population stratification），最好改用基于家系资料的关联分析，如 TDT 或 ASP 法（参见 8.5.3 节）。

9.2.1.3 候选基因的研究

关键在于候选基因的选择和该基因中 SNP 的选择。分析结果为阳性时，仍要注意所选基因是否具有生物学意义，以及所选 SNP 是否与该基因的功能改变有关。一项有意义的研究结果最好能用不同方法重复其阳性结果。例如，用病例对照分析的结果，最好能用 TDT 或 ASP 方法重复出来。

9.2.1.4 极端不一致表型分析（extreme discordant phenotype，EDP）

该方法属于病例对照分析用于药物反应性研究的特殊方法[2]。一般地，随着药物剂量的增加，更多的患者表现出毒性效应，即毒性发生率增加。同时人群中仍然使用该药而不表现毒性的人越来越少。将那些接受最小剂量但表现出强烈毒性反应者归为“极端敏感型”

组，另一些接受最高剂量但不出现毒性反应者归为“极端耐受型”组。以这两组分别作为病例和对照组，分析暴露于候选基因的 SNP 或等位基因下的相对危险度或几率比值，来揭示药物反应型与易感基因的关联性。这项研究中，不考虑那些给予中等剂量或出现中度反应的人，因此大部分研究对象未被考虑，为了使样品量达到一定要求，往往所研究的群体要足够大。EDP 方法是解析药物反应性的常用方法，在此基础上，结合 ASP 方法发展为修正 EDP 法。

9.2.1.5 样品量估计

基于 SNP 的病例对照研究中，所需样品量与 SNP 等位基因频率和要求的统计效能(statistical power)有关。表 9－3 所列为几率比值超过 1.5，统计效能等于 80%，Ⅰ类错误概率 α 值分别为 0.05 和 0.005 时所需样本量[1]。

表 9－3 SNP 病例对照研究样本量的估计(OR≥1.5，统计效能 =80%)

等位基因频率%(对照组)	显性方式			隐性方式		
	病例组流行率(%)	样本量		病例组流行率(%)	样本量	
		$\alpha=0.05$	$\alpha=0.005$		$\alpha=0.05$	$\alpha=0.005$
10	19	430	711	1	6 113	10 070
20	36	311	516	4	1 600	2 637
30	51	308	512	9	769	1269
40	64	354	590	16	485	802
50	75	456	762	25	363	602
60	84	661	1 107	36	311	516

9.2.2 全基因组连锁不平衡作图[1]

进行基因组 SNP 连锁不平衡分析的假设是 SNP 之间存在 LD(即 SNP 单体型)，因此该方法的分析效能取决于 SNP 的等位基因频率和这些 SNP 之间的 LD 的区间大小。群体混杂、遗传漂变、基因突变、自然选择等多种因素影响 LD 区间大小，平均为 5～150 kb，参见 3.5.2节。一种简化全基因组扫描的方法是专门针对位于人类基因组中约 3 万个功能基因上的 LD，这可避开大量位于基因间序列的 LD，从而减少工作量，提高检测效率。

鉴定 LD 的方法主要是通过实验，进行 SNP 的基因分型。所涉及的分析方法常用最大似然性分析，利用 EM 运算法则，计算群体中 LD 的频率。LD 作图分析中，由于所采用的 LD 分析不够精确，仍可遗漏某些关键的 SNP，导致分析结果的假阴性。补救的办法是在锁定某个候选基因后，最好对该基因内所有的 SNP 进行测定和统计分析。

3.5.2 节中详细论述了单核苷酸多态性(SNP)的连锁不平衡现象及 LD 区间大小的影响因素。LD 的发生与群体混合、自然选择、遗传漂变、新突变产生等因素有关。根据 LD 的分布进行疾病基因的关联分析是目前多基因性状研究的三大常用方法之一(其他两种方法为 TDT 和 ASP)，简称 LD 基因定位(LD mapping)。影响 LD 基因定位的因素很多，包括标记物的密度、样品大小、疾病发病率、等位基因频率、人群和种族差异等。用于 LD 分析的样

品，特别是病例对照研究样品，应作 Hardy－Weinberg 平衡检验。LD 基因定位分析步骤为：①随机群体中进行 SNP 作图；②SNP 单体型分析；③SNP 单体型中标签 SNP 的鉴定；④分析样品的标签 SNP 基因型分析；⑤关联分析和统计推断。LD 的表示是通过计算 D'值。D'的计算可通过软件给出，如 GOLD 软件，参见 http://www.sph.umich.edu/csg/abecasis/GOLD/。

用于 LD 作图样品量的要求，需要根据 SNP 等位基因频率、受检 SNP 的数量、LD 的大小来确定。随着连锁不平衡性的 D'值的降低，可检测的相对风险也降低，关联性研究越难判断，因而需要的样品量更大。样品量与 D'值的关系如下：

$$N/(D')^2 \qquad \text{（公式 9－2）}$$

上式中 N 表示当 D'值等于 1 时（完全 LD），关联分析所需样品数量。如进行某个药物遗传性状 SNP 的关联分析所需样品为 400（参见表 9－3），当另一个 SNP 与该药物遗传性状 SNP 的 D'值为 0.7 时，则进行关联分析所需样品数为 816。

9.2.3　药物基因组学研究方法的设计[1]

药物基因组学研究大致分为干涉性研究（interventional study）和观察性研究（observational）两大类。前者又称临床试验（clinical trial），后者主要包括病例对照研究（case－control study）和队列研究（cohort study）。临床试验的优点是能对许多试验影响因素通过对照组设计进行消除，如治疗适应证、年龄、性别、剂量、给药时间和顺应性等。同时临床试验中，通过服用安慰剂（placebo）作为对照，有利于揭示药物与基因间相互作用的内在联系，并区分疾病基因和药物遗传性状。随着人类基因组计划的完成，临床试验的设计可针对整个基因组的表达，这是通过比较药物对不同组织或细胞株基因表达谱的区别，寻找药物的靶向基因，从而鉴定药物反应性的方法。基因表达芯片或药物筛选芯片应运而生，参见第 7 章。

药物上市前，如果未进行Ⅳ期临床试验，常须采用病例对照研究和队列研究进行药物疗效和安全性观察。病例对照研究主要采用回顾性分析，记录服药并出现药物反应者（病例）和服药但未出现药物反应者（对照），根据相应的基因型，通过计算几率比值，分析基因型与药物反应的关联性。队列研究主要采用前瞻性分析，观察一组服药后的队列，哪些出现药物反应，哪些未出现药物反应，再比较这两组中与某种基因型出现的频率，即相对危险度，以分析基因型与药物反应的关联性。通常需要根据研究的性质来决定采用何种观察性研究方法。队列研究，一般需要观察 5 000 例以上患者，以保证药物反应组有 250 例以上，但通常需要连续观察较长时间，才能评价药物反应是否出现。病例对照研究较适合于发生率低于 5% 的药物反应性状，以及对药物的长期（10～20 年）有效性的观察。表 9－4 比较了药物基因组学研究中 3 种常用设计方法的优缺点。

表9-4 药物基因组学研究的常用设计方法比较

	病例对照研究	队列研究	临床试验
设计原理	病例:服药并出现药物反应者; 对照:服药但未出现药物反应者	选择一队患者进行给药(他们服药前不知道是否对该药物有反应),在一定时间内按是否出现反应进行分组	随机对患者进行分组:服药(或者安慰剂)组、不服药组
药物反应发生时间	发生在选择研究对象之前,属回顾性研究(需要界定病例和对照的标准)	发生在选择研究对象之后,通常是前瞻性研究,也可以是回顾性研究(根据医疗记录)	发生在选择研究对象之后,属前瞻性研究
观测指标	基因与药物的相互作用,但无法测定基因或药物的单独作用	基因与药物的相互作用,但无法测定基因或药物的单独作用	基因或药物的单独作用以及基因与药物的相互作用
主要优点	对罕见效应最有效(如药物性肝炎)	不是根据是否对药物出现反应而选择研究对象,因此较适合连续性药物反应性状	最准确,可排除许多影响因素的干扰
主要缺点	剂量、适应征、给药时间等因素未标准化,因此可对结果分析产生干扰	对前瞻性研究,其缺点同临床试验;对于回顾性研究,其缺点同病例对照研究	需要足够的随访时间以观察药物效应的出现

9.2.4 药物基因组学常用技术平台[1,4]

药物基因组学研究中,涉及的方法有多种,主要为三大技术平台:SNP 分型、基因表达芯片和生物信息学。这里着重介绍 SNP 分型技术平台。基因表达研究可用于候选基因的筛选,在药物筛选中也起重要作用。目前主要应用 cDNA 芯片技术,其他如差异显示(DD)、代表性差异分析(RDA)、抑制性消减杂交(SSH)、基因寻呼(GeneCalling)等也是基因表达研究中的常用方法。有关基因表达芯片内容参见第7章。药物基因组学中涉及的生物信息学内容主要是数据管理、DNA 序列分析、SNP 数据库、蛋白结构分析与预测、蛋白与蛋白或配体之间的相互作用、代谢途径分析、表达数据库、生物影像学、数据统计分析等,相关内容分布于本书不同章节。其他药物基因组学研究相关的技术平台还包括基因组学、蛋白组学、药动学、药效学等。

SNP 的测定主要基于 PCR 技术基础上,与传统点突变检测方法相似,可大致分 SNP 筛查方法和 SNP 分型方法两大类。前者主要是针对某基因区段中的未知 SNP 的检测方法,如单链构象多态性分析(SSCP)、变性高效液相色谱(dHPLC)、变性梯度凝胶电泳(DGGE)、异源双链分析(HA)等;后者则是在已知的与药物反应性有关的 SNP 基础上,对患者进行 SNP

分型方法，如限制性片段长度多态性分析（RFLP）、引物延伸法（PEX）或焦磷酸测序技术（pyrosequencing）、等位基因特异性寡核苷酸杂交（ASO）、等位基因特异性扩增技术（ASA）以及 DNA 直接测序技术（DS）等[4]。适合于药物基因组研究中的 SNP 分型方法，要求操作简便，准确性高，高通量，成本低。目前基于上述点突变检测方法原理，已发展出几种常用的自动化、高通量的 SNP 检测技术，参见表 9－5。各实验室可根据自己的经济条件、技术水平、检测目的、检测要求等合理选用。

表 9－5　常用 SNP 检测方法的比较

检测技术	检测原理	检测通量	检测成本	准确性	自动化程度	操作简便性
dHPLC 仪	单链构象 dHPLC	较高	较高	较高	一般	中等
质谱仪	质量 电荷	高	极高，多重 PCR 时↓	极高	高	中等
焦磷酸测序仪	引物延伸 焦磷酸化	高	中等	高	高	高
荧光定量 PCR 仪	ASO 荧光探针	高	较高	较高	高	高
DNA 芯片	ASO 微阵列	高 SNP 数 低样品数	极高	低	低	极低
DNA 测序	Sanger 法 毛细管电泳	低到中	高	高	高	高

9.2.5　个体化医学在诊断和治疗中的应用[1,5]

在 3.4.2 节有关应用基因组学发展中已介绍药物反应性在个体化医学中的应用。个体化医学的发展主要建立在人类基因组多样性和药物基因组学的基础上。目前已明确许多多因子性状的易感基因或位点，参见 3.4.1 节表 3－3。同时，药物遗传性状也得到进一步阐明。表 9－6 所列为常见药物反应的相关基因。这些多基因性状和药物反应性状已逐渐应用于临床诊断和指导个体化治疗，大大丰富了预测医学（主要是指“差异诊断”，differential diagnosis）和基因诊断的内容，也使药物治疗更加趋于安全和有效。药物基因组学在个体化医学中的应用目标是期望每个个体达到用药的最佳效果，同时副作用降到最低，即为不同的基因型提供最适合的药物（“基因型适应证”），或者确定“基因型禁忌证”。

目前临床上已开始对治疗前的患者进行基因型分析，以避免那些可能产生的严重药物副作用。如利用抗癌药 5－氟尿嘧啶治疗前，需要测定二氢嘧啶脱氢酶基因的基因型，以避免产生严重的毒性。器官移植中，利用免疫抑制剂咪唑硫嘌呤治疗前，须筛查硫代嘌呤 S－甲基转移酶的基因是否因突变失活，以避免致死性的血液系统毒性作用。

最近 Chang 等[5]报道用基因表达芯片方法预测 docetaxel 对乳腺癌治疗反应性，这是首次利用基因表达图谱来预测药物个体反应性，比单纯利用某个基因的 SNP 来评价可靠性更高。该项研究中，共研究了 24 例乳腺癌患者，发现 11 例（46%）对 docetaxel 敏感，13 例

(54%)表现为耐受。该研究发现92种基因的表达在这两组病人中出现的表达量明显不同($P=0.001$)。敏感型患者中,78种过度表达的基因主要涉及细胞凋亡、细胞骨架、细胞黏附、蛋白转运、信号传导、RNA剪切或转运等;而耐受型患者中,14种过度表达的基因主要与蛋白质翻译、细胞周期和RNA转录有关。分别以这两组基因的变化进行药物反应性的评价,阳性和阴性预告值分别为92%和83%,ROC曲线下面积为0.96,具有良好的预测作用。

表9-6 常见影响药物反应的基因型

药物	靶基因	基因型	药物反应性
HERCEPTION (trastuzamab)	HER2	表达型	该药仅对约1/3的乳腺癌表达该基因的患者有效
咪唑硫嘌呤 (azathioprine) 6-巯基嘌呤	TPMT	零突变 (null mutation)	突变后导致该酶活性丧失,缺乏对该药的降解,而易产生对造血系统的致命性毒性
5-氟尿嘧啶 (5-FU)	DPD	零突变	突变后导致该酶活性丧失,缺乏对该药的降解,产生毒性
肼苯哒嗪 异烟肼 磺胺药	NAT2	零突变	突变后导致该酶活性丧失,产生毒性或过敏反应
氯丙嗪 氟哌啶醇 可待因 等	CYP2D6	sparteine-debrisoquin 等位基因	药物治疗无效(无反应)
ABT-761 (抗哮喘药)	ALOX5	启动子SNP Sp1结合基序	药物治疗无效(无反应)
抗AIDS药物	CCR5	Δ32	具有该基因型者对抗AIDS药物无效
地高辛等	MDR1	3435C>T	引起血浆地高辛浓度增加
华法林 水杨酸盐 安定 等	ALB	Lys313Asn As p365His	与血清白蛋白结合力降低,转运障碍,影响药动学
STI-571	BCR-ABL	Thy315Ile	出现耐药性

基因芯片在药物基因组学的临床应用的另一大领域是用药方案的选用和预后评价。有报道对几种常用的急性淋巴细胞性白血病化疗药物的疗效观察中,通过研究不同临床分期患者用药后基因组水平的变化,以及比较不同化疗方案对基因组表达的影响,进行聚类分析(clustering)。根据聚类资料既可评估疗效和预后(复发或消退),也可评价化疗药物联用与单用效果的区别,用于指导用药方案的确定。总之,基因芯片结合药物遗传学或药物基因组学的研究,在个体化治疗中的应用尚处于初步应用阶段,预计随着技术的完善和成本的降

低，其应用将更加广泛和普及。值得注意的是，人体是一个极其复杂的生物体，个体化诊断只是提供一种可能性的推断，但并非是绝然性的，还受许多环境因素或基因间相互作用的影响。

9.3　药物筛选与计算机辅助药物设计

9.3.1　药物研发的基本过程[1,6]

新药发现和研究是一个耗费巨大（在美国平均每个药品上市约需 8 亿美元）、漫长（平均 10～12 年）的过程。目前世界各大制药企业正在试图利用新技术来减少新药开发的费用和时间，其中药物基因组学和化学信息学是这些新技术中的代表。图 9－3 示新药研发的基本流程及相关的药物基因组研究。新药研发分以下几个阶段：

研发阶段		化合物数目	受试人数	目标	药物基因组研究内容
临床前期（3～12年）	靶分子鉴定			鉴定靶分子	发现或鉴定疾病基因，用做新药靶分子
	高通量筛选	10 000		鉴定前导化合物	筛选影响药物结合的靶分子编码区SNP
	化合物的优选（0期）	250	动物	筛选候选药物	筛选影响药物结合的靶分子编码区SNP
临床前（4～8年）	Ⅰ期	5	15～40	研究在人体中的短期毒性和药动学	样品数太少，不进行药物基因组评价
	Ⅱ期	3	100～300	Ⅱa:在受试人群中进行有效性、药动学、副作用的先期评价 Ⅱb:核心受试者双盲法与安慰剂评价有效性	利用已鉴定的影响靶分子性的SNP评价药物反应性，
	Ⅲ期	2	1 000～5 000	再次对用药1年以上的核心受试者进行双盲试验	鉴定药物反应性的SNP，预测副作用，寻找替代新药
上市后期	Ⅳ期	1	群体	投放到市场，药物警惕性观察	基因型适应证或禁忌证
	新适应证或副作用	1	群体	药物相互作用，老药新用或撤市	同Ⅳ期

图 9－3　药物研发过程及相关药物基因组研究

9.3.1.1　靶分子的鉴定

目前面市的药物靶分子共 500 种左右，大致分为 3 类：酶、受体、循环中的蛋白分子，所占比例分别为 45%，28% 和 11%。随着人类基因组计划的完成，新基因的不断阐明，为药物

研发提供了越来越多的靶分子。对于某些具有多种效应分子的药物，利用药物基因组技术在基因组水平进行研究，有助于全面认识药物的作用机制。

9.3.1.2 临床前期研发

前导化合物(lead compound)的鉴定是新药研发中的关键。所谓前导化合物是在众多类似药物中筛选出的非常具有开发潜力的化合物。这需要建立在全面的药理学(药动学)和毒理学研究的基础上。此阶段药物基因组学的研究，重点针对靶基因编码区的 SNP，剔除那些与常见的影响药物反应性的代谢酶或受体多态性具有相互作用的化合物，同时在基因组水平发现药物的副作用。临床前期主要是实验室研究和动物实验。临床前期完成后，需向主管部门申请新药研究试验。

9.3.1.3 投放市场前的临床试验Ⅰ、Ⅱ、Ⅲ期

临床Ⅰ期试验主要是初步在健康志愿者中对候选药物进行药动学和毒性试验。临床Ⅱ期又分Ⅱa 和Ⅱb，Ⅱa 期主要是先期评价药物的有效性，并确定药物的平均剂量、用药次数、副作用等。通常分 2～3 种剂量，将病人分组研究，每组约 50 人。此期可同时进行靶分子的基因分型研究，观察药物遗传性状。Ⅱb 期主要是有效性的确认试验。通常采用双盲法，加安慰剂作对照。同时根据受试人员数量，选择核心受试组(pivotal trial)。Ⅲ期临床试验主要是在大规模人群中有效性和安全性的评价，通常采用随机分组的双盲试验，同时评价药物之间的相互作用。资料分析时，除按表型分组比较外，还需按基因型分组比较，并尽量考虑各种混杂因素(confounder)和协同变量(co－variable)的影响。

9.3.1.4 投放市场后的Ⅳ期

药物经过临床试验后，向药品管理局申报上市，详细阐明药物的有效性和安全性。Ⅳ期研究是在药物投放市场后，继续对药物进行评估，主要针对药物的成本收益分析(药物流行病学)，同时基于群体的药物遗传学研究，有助于进一步阐明药物反应性与基因多态性之间的关系。此期也需要进一步观察药物之间的相互作用，并利用关联分析，通过长期观察，分析与药物副作用的关系或发现药物的其他药理作用。一旦发现药物新的严重毒性作用，则需要考虑药物撤出市场。

新药研发中的药物基因组学研究，也涉及到许多伦理、道德和法律问题。如上述过程中涉及的动物实验和人体试验必须经过各类 IRB(Institutional Review Board)机构的认可。临床试验中所有志愿者必须填写知情同意书，涉及基因型分析和有关疾病信息，研究人员必须遵守保密原则，并充分尊重志愿者的自主权和隐私权。特别注意避免“基因歧视”现象和利用基因型分析用于医疗保险凭证。各个国家对于新药审批都有一套严格的法律法规制度，没有经过审批的药物应视为非法药物。

前面已经提到，新药研发的关键是发现前导化合物。前导化合物的发现方法有多种。例如：①偶然凭运气发现，如青霉素的发现。②天然资源中的寻找，如大多数中草药的发现。③通过研究靶分子——代谢酶的底物、抑制剂、活性中心等，发现潜在的药物。如大多数底物类似物、竞争性抑制剂等的发现。④对类似疾病，试用相同药物。如革兰阳性菌感染药物。⑤老药新用的发现，如阿司匹林的抗血栓形成作用、“伟哥”的性兴奋作用的发现等。⑥药物基因组筛查：如利用多种肿瘤细胞株和表达芯片技术进行的肿瘤药物的筛选。⑦药物锚定：利用化学信息学方法，建立药物与靶分子相互作用的模型，进行大规模筛查，确定药

物的靶向分子。⑧计算机辅助药物设计：基于靶分子原始性设计一种可与之相互作用的药物。

本章主要讨论涉及化学信息学内容的最后两种途径。

9.3.2　药物锚定方法[1,7-9]

传统的药物研发中，主要依靠对天然存在的生物活性物质的化学修饰，或者利用“化合物库”进行高通量筛选，得到前导化合物。随着人类基因组数据的不断增加，药物靶分子数目不断增多，这些方法的使用则受到一定限制。近年来，随着结构基因组学的发展（见第6章），许多蛋白质的三维结构得到阐明，同时其配体结合部位也已清楚，一些蛋白复合体中蛋白亚基之间的相互作用也已明确。基于这些信息，可将某些化学小分子锚定在靶向蛋白质的结合部位，称为药物锚定（drug anchoring 或 drug docking）。这是目前新药研发中必不可少的工具。

在筛选前导化合物过程中，需要对候选化合物进行系统研究。往往侧重候选化合物相关的或类似的化合物的研究，通过对各自结构和理化特征的分析，比较其特异性和药理作用，并借助计算机模拟药物与靶分子的相互作用和统计模型的建立，进行化合物筛选。因此可以看出，药物锚定中的关键化学信息学方法是前导化合物类似物的筛选、药物与靶分子相互作用预测以及多元回归分析。

（1）前导化合物类似物的筛选主要借助已有的化学数据库如 World Drug Index、LIGAND 或 MDL 化合物目录（http://www.mdli.com），对其中的化合物进行分类筛选。常用的筛选方法有[7]：聚类法（cluster - based）、分散法（partition - based）、不相似性法（dissimilarity - based）、优选法（optimization - based）。聚类法和分散法分别根据化合物分子结构或物理化学特征进行分组，在每组中挑选一个代表性药物进行分析。不相似性法则相反，主要按照结构上的完全不同性原则选择化合物。优选法也是在对化合物差别性定量分析基础上，进行筛选。

（2）筛选后的化合物进一步利用化合物结构图形，比较同类化合物的结构异同点，预测其空间结构和物理化学特性，以评价筛选的效果。预测物理化学特性的方法也是目前化学信息学的重要内容[8]。这主要针对药物的结构特点预测其药动学特点，即通常所谓的 ADME 特性[吸收（absorption）、分布（distribution）、代谢（metabolism）、排泄（excretion）]。其中主要是对药物的吸收、渗透性（permeability）、可溶性（solubility）等进行预测。

（3）在获得其结构信息后，利用前导化合物与靶分子相互作用的实验数据，建立相互作用模型，比较其类似物相互作用的方式，从中筛选理想的化合物。

（4）在取得相关实验数据基础上，对分子特征和药理作用的关系进行分析。这是20世纪60年代由 C. Hansch 等建立的定量结构与活性相关性分析（quantitative structure - activity relationship，QSAR）方法。QSAR 分析还与药效学或药动学研究相结合，以进一步分析结构类似物的药理效应。例如在可的松药物的研究中，利用 QSPR 分析结果整合到皮质类固醇受体的药动学模型，来分析脱氢皮质（甾）醇类似化合物的药理作用[9]。

9.3.3 计算机辅助药物设计[1,7,8]

计算机辅助药物设计(computer - aided drug design,CADD)是根据靶分子已有的信息原始性(*in silico*)设计与靶分子相互作用的小分子化合物,而药物锚定是将已知的小分子化合物,借助计算机分析,寻找其相互作用的靶分子,两者的研究方向相反,但研究目的相同,都在于发现前导化合物。形象地说,CADD 如同手上有一把待开启的锁,需要锁匠配一把能打开的钥匙(俗称"配钥匙")。药物锚定好比宾馆服务员手上有一把钥匙,但不知是开哪个房间的,需要找到可打开的门锁(俗称"找门锁")。两者的比较见图 9 - 4。以下阐述 CADD 的一些要点。

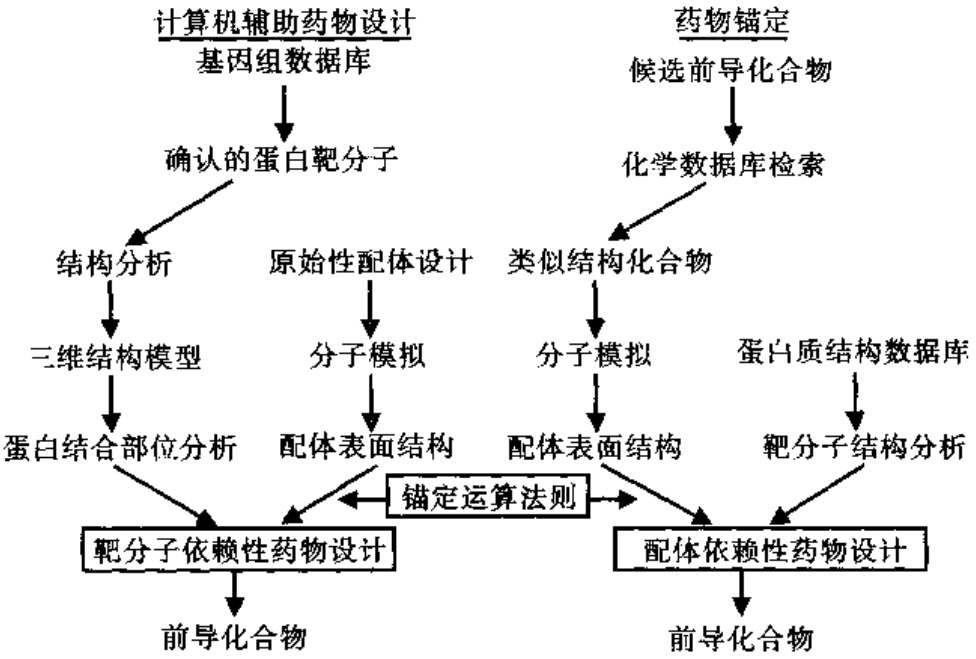


图 9 - 4 计算机辅助药物设计与药物锚定的比较

(1)利用 NMR 或 X - 衍射方法已阐明了许多蛋白质的空间结构,并储存在 PDB 数据库中(http://www.expasy.ch/spdbv)。目前 PDB 已有超过 1.5 万个条目,其中不少属于冗余的(redundancy),参见 6.8.1 节。数据库中尚未收录的靶分子的结构,可以根据其同源蛋白质的结构来设计,当然,所得到的化合物结构的准确性也受同源程度影响。

(2)由于 SNP 或突变的存在,可影响蛋白质的结构,合理的药物设计应该尽量考虑这种多态性对药物与靶分子结合的影响。

(3)目前依靠实验得到的蛋白质结构仍然有限,计算机辅助结构预测(computer - aided structure prediction, CASP)成为蛋白质结构分析的重要手段(参见第 6 章),并用于 CADD。

(4)靶分子与配体的相互作用方式有多种,包括氢键、离子键、疏水键等,有效的相互作用取决于配体与靶分子之间空间结构的互补性相互作用,预测时需要建立准确的锚定运算法则(docking algorithm)。设计的药物分子与靶分子的亲和力由自由能(free energy)决定,自由能的变化可根据焓(enthalpy)和熵(entropy)的变化来推算。

(5)配体从头开始(*ab initio*)设计遵循随机原则,同时必须符合化学规则。这种装配方

式通常是先根据不同结合位点数据，再进行整合，获得一个拼装分子，再重新校验完整分子的物理化学特点后，进行模拟锚定分析。

（李　巍）

参 考 文 献

1　Licinio J, Wong ML. Pharmacogenomics: The Search for Individualized Therapies. Weinheim: Wiley - VCH, 2002. 1 ~ 157

2　Rimoin DL, Connor JM, Pyeritz RE, *et al*. Emery and Rimoin's Principles and Practice of Medical Genetics. 4th ed. London: Churchill Livingstone, 2002. 590 ~ 627

3　Ring HZ, Kroetz DL. Candidate gene approach for pharmacogenetic studies. Pharmacogenomics, 2002, 3: 47

4　陆国辉，李巍，黄艳仪，等. 产前遗传病诊断. 广州：广东科技出版社，2002. 198 ~ 219

5　Chang JC, Wooten EC, Tsimelzon A, *et al*. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet, 2003, 362: 362

6　Lesk AM. Introduction to Bioinformatics. Oxford: Oxford University Press, 2002, 216 ~ 272

7　Willett P. Chemoinformatics - similarity and diversity in chemical libraries. Curr Opin Biotech, 2000, 11: 85

8　Blake JF. Chemoinformatics - predicting the physicochemical properties of drug - like molecules. Curr Opin Biotech, 2000, 11: 104

9　Mager DE, Pyszczynski NA, Jusko WJ. Integrated QSPR - pharmacodynamic model of genomic effects of several corticosteroids. J Pharm Sci, 2002, 92: 881

第10章　统计生物信息学

统计生物信息学(statistical bioinformatics)是生物信息学与生物统计学(biostatistics)的一门交叉学科,主要利用统计学的基本理论和方法解决生物信息学中涉及的有关数理统计问题,如概率的计算、统计遗传学、遗传流行病学、药物基因组学、生物芯片数据分析等。实际应用中,在掌握基本的资料分布基础上,首先提出检验假设;然后将数据资料进行合理分类和整理,建立数据文件;除非是很简单的数据分析用手工计算外,基本上是利用统计软件,根据研究者的实验设计和研究目的,选用合适的统计方法进行分析,或者根据研究目的,建立统计分析模型再运算;最后也是最关键的步骤是对统计结果进行合理的解释。本章仅介绍概率论与统计分析的基本方法,对于一些较高级的统计方法,如判别分析(最大似然法、贝叶斯法)、聚类分析、主成分分析等,在第7章有关基因芯片数据分析和第8章统计遗传方法中有所涉及,这里不再赘述。此外,统计生物信息学中的常用统计表,如标准正态分布曲线下的面积、t 界值表、χ^2 界值表和 F 界值表等因限于篇幅,本书未列出,需要时请读者查阅相关的统计学参考书。

10.1　概率论基础[1]

概率(probability,P)是用来表示事件 A(如某种遗传病)发生的可能性,即在 n 次重复性实验中,发生 A 的次数 n_A 占总数的比例,用公式表示为:

$$P(A) = \frac{n_A}{n}$$

其数值在0~1之间。当概率为0时,表示该事件不可能发生;当概率为1时,表示该事件必然发生。概率的表示方式可以是小数(如

0.5)、分数(如1/2)、百分数(如50%)或比数(如1:1)。当概率为0.25时,表示发生的可能性为1/4(或25%),或者说3/4(即75%)的可能性不发生。概率计算中有几个基本法则,如加法法则、乘法法则和互补原则。

(1)加法法则(law of addition) 如果两个或两个以上的事件是相互排斥的,事件A的发生概率为P_1,事件B的发生概率为P_2,则出现A或B的概率为$P_1 + P_2$。例如,怀孕后是男孩的概率为0.5,是女孩的概率也为0.5,不可能同时为男和女,因此胎儿是男或女的概率等于1。

(2)乘法法则(law of mutiplication) 当两个或两个以上事件的结果是独立存在或可以同时发生时,事件A的发生概率为P_1,事件B的发生概率为P_2,则发生A和B的概率为$P_1 \times P_2$。例如一对夫妇生第一个小孩为男孩的概率为1/2,生第二个小孩为男孩的概率也是1/2,那么该夫妇生育两个小孩都是男性的概率为$1/2 \times 1/2 = 1/4$。

(3)互补原则(rule of complement) 如果某事件E发生的概率为$P(E)$,则不发生的概率为$P(\bar{E})$,则$P(E) + P(\bar{E}) = 1$,只要知道其中之一,便可推算出另一个概率。

条件概率(conditional probability)是概率论中常用的一个概念,是指某样本空间(sample space)中的两个事件E和F,在E发生的前提下,再发生F的概率,记作$P(F|E)$。条件概率的概念是著名的贝叶斯(Bayes)定理的分析基础,分别在8.4.1和8.7.1节中已作描述,这是统计分析和遗传风险分析中的常用算法。

10.1.1 随机变量

变量(variable)是指某种可测量的用于区别不同物体的特征或特性。当我们测量某种特征时,通常建立一个数字模型进行表达。例如,某种特定组织中某个基因mRNA的表达量就是一个变量。数值反映转录水平的强度。一定条件下产生相同表达量mRNA的基因赋值相同。

统计学变量主要分两类:定性或定量变量。定量变量进一步可区分为连续变量(continuous variable)和离散变量(discrete variable)。连续随机变量可以是某区间内任何实数。例如,基因表达水平便是一个连续随机变量。由连续随机变量所得到的数据称为连续性数据。离散变量通常是测量或计数中可定的整数值。例如DNA微阵列上基因的数目就是一个离散变量,其取值为正整数。随机离散变量所代表的数据称为离散数据。

10.1.2 概率分布函数和频数分布[1,2]

概率分布(probability distribution)是用来描述随机变量与概率发生之间的关系,通常用表格、图形或函数来表示。概率分布主要分离散变量分布和连续变量分布两大类。常见的离散分布包括超几何分布(hypergeometric distribution)、二项分布(binomial distribution)、泊松分布(Poisson distribution)、帕斯卡分布(Pascal distribution)等。常见的连续分布包括β分布(beta distribution)、γ分布(gamma distribution)、χ^2分布(chi-square distribution)、t-分布(t distribution)、F-分布(F distribution)和正态分布(normal distribution)。统计学分析中,首先是要确定数据的分布类型,然后进行参数估计和假设检验。这里将不对各分布函数及其主要参数,如均数和方差的计算展开讨论。统计方法中涉及有关分布时再分别论述。有

关标准正态分布(即 Z 分布)曲线下面积(概率)的计算方法参见 7.4.2 节。

在计数资料分析中,常涉及频数分布的概念。它是用来统计每个变量值所处的区域,从而将资料进行收集整理的方法。表 10 - 1 是 100 个基因用 cDNA 微阵列检测所得到的结果。初步看来,很难得到这些数据的简明结果。需要进一步加以整理成更紧凑的形式并能说明一些问题。这便是频数分布的由来。

表 10 - 1 cDNA 微阵列测定 100 个基因表达的强度

Gene	Intensity	Gene	Intensity
gene1	83.7	gene36	5.8
gene2	11.1	gene37	259.8
gene3	327.6	gene38	110.4
gene4	86.2	gene39	60.9
gene5	33	gene40	100.3
gene6	175.1	gene41	33.8
gene7	110.6	gene42	238.5
gene8	652.4	gene43	52.3
gene9	363.3	gene44	277.6
gene10	16	gene45	368.3
gene11	975.7	gene46	419.9
gene12	748.3	gene47	12.2
gene13	98	gene48	206
gene14	28	gene49	20
gene15	155.8	gene50	129.4
gene16	73.2	gene51	90.7
gene17	44.9	gene52	596.7
gene18	32.7	gene53	331.3
gene19	183	gene54	52.6
gene20	144.6	gene55	141.3
gene21	29.1	gene56	170.3
gene22	53	gene57	9.8
gene23	582.1	gene58	669.7
gene24	75.7	gene59	251.3
gene25	442.6	gene60	200.3
gene26	11.7	gene61	31.8
gene27	320.2	gene62	887.6
gene28	428.9	gene63	54
gene29	363.7	gene64	24
gene30	603.4	gene65	192.4
gene31	72.2	gene66	961.8
gene32	34	gene67	445.6
gene33	471.5	gene68	152.1
gene34	326.6	gene69	7.9
gene35	28.3	gene70	889.4

续表 10－1

Gene	Intensity	Gene	Intensity
gene71	713.4	gene86	17.2
gene72	718.7	gene87	474.9
gene73	549.8	gene88	43.3
gene74	296.3	gene89	394.6
gene75	331.9	gene90	369.6
gene76	105.2	gene91	373
gene77	96.7	gene92	17.8
gene78	703.5	gene93	305.2
gene79	84.9	gene94	602.4
gene80	598.2	gene95	516.9
gene81	368.9	gene96	39.7
gene82	99.4	gene97	115.8
gene83	155.2	gene98	702.2
gene84	61.4	gene99	159.9
gene85	603.1	gene100	25

连续资料的频数分布可以用不同的方式来表示，如频数表、直方图(histogram)等。这种图表的制作通常是由计算机来完成。用户必须清楚了解其原理并懂得如何解析其结果。表 10－2 是根据表 10－1 绘制的频数分布表。间隔宽度设置为 100，读者不难理解表中各列的计算方法和所代表的意义。经过表 10－2 的处理，原始数据得到压缩而且可说明一些问题。如果观察第(1)和第(4)列，可以很容易发现这 100 个基因所聚集的强度范围。第(6)列的累积相对频率中，可看出某一栏目上下观察数目的多少，如超过半数的基因的表达强度在 200 以下。

表 10－2　根据表 10－1 基因表达数据绘制的频数分布表

(1) 间隔	(2) 中点	(3) 频数	(4) 相对频率	(5) 累积极频率	(6) 累积相对频率
1～100	50	40	40%	40	40%
101～200	150	16	16%	56	56%
201～300	250	6	6%	62	62%
301～400	350	13	13%	75	75%
401～500	450	6	6%	81	81%
501～600	550	5	5%	86	86%
601～700	650	5	5%	91	91%
701～800	750	5	5%	96	96%
801～900	850	2	2%	98	98%
901～1 000	950	2	2%	100	100%

用作图法来表示频数分布信息量很大。例如将表 10－1 和表 10－2 中的数据用直方图

表示，见图 10－1。横轴表示基因表达强度，纵轴表示以 100 为间隔宽度统计所得频数。这种图形较表 10－2 直观。

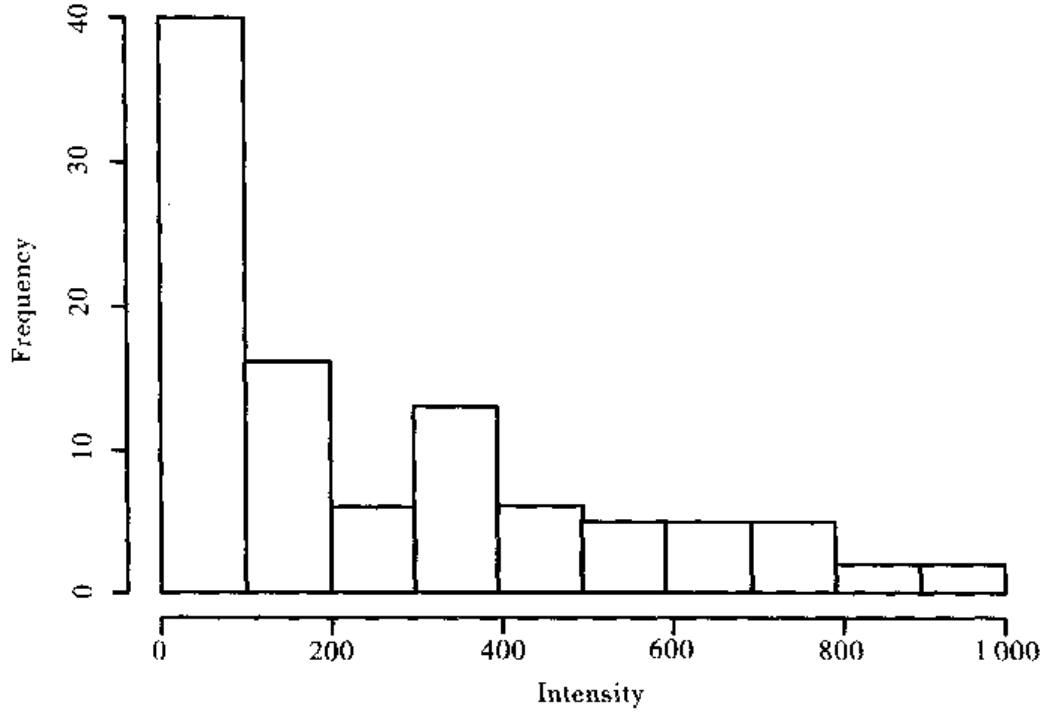


图 10－1 表 10－1 数据的直方案

10.1.3 总体与抽样分布

生物学研究中，常面临不同的研究对象，这个研究对象所代表的全部信息，称为一个总体(population)。许多生物信息处理中，我们不可能对总体中所有个体逐个检测，一般是随机抽取总体中的一些样本(sample)进行检查，又称抽样(sampling)。上面有关频数分布主要描述资料的集中趋势。在推论性统计学中，需要通过从总体中抽样得到的资料进行推断。这种过程常会有一定程度的误差，又称抽样误差(sampling error)。统计分析中，必须了解这种推论中的确定性和不确定性因素，这仍然涉及概率的问题。

我们先来看一个 DNA 序列分析的例子。假如我们有下列两种分别来自不同物种的序列，下划线代表相同核苷酸。

G T C T G T G G G C A A T A C G G A T C C G T C A A

G A C A T T A T G G A T T T C T G T A G C G A T A T

现在我们想知道这两种序列是否具有明显的相似性，即预测它们是否具有共同的祖先。如果这两个序列是由 4 种碱基(A，T，G，C)随机得到的序列，其相似性应该接近 1/4。上述例子 26 个碱基中，我们可预期得到 6 个相同的碱基，但事实上已经发现 11 个相同碱基。这种结果是否由于随机产生的？其可能性需要用概率原理来分析。如果因随机产生的序列的概率很低，可认为其他因素在起作用，即两者可能来自同一共同祖先。

抽样分布(sampling distribution)属于概率分布的范畴，是通过对抽样所得到的样本的分布进行表述的方式。统计分析中，需要推断这种抽样分布的基本参数，如均数、标准差和置信限等(详见 10.2 节)，从而判断抽样误差的大小，得到这种抽样所得到的样本是否完全来

源于同一总体或总体的结论。

10.2 统计学基本概念

上节提到的有关总体、样本、变量、频数分布和抽样分布的概念是统计学的基本概念。其中，频数分布将数据简化成更紧凑的形式，是描述性统计学的一种，用于描述数据分布的特征和提供一种简单的总结数据，阐明典型观察值分布和集中趋势。同时也反映数据分布的变异性和分散性。描述性统计学中常用均数和方差来分别代表集中趋势和离散趋势的指标，又被称为点值估计(point estimation)。置信区间(confidence interval)是区间估计(interval estimation)中的主要指标，主要利用抽样分布估计总体均数的概率范围。一般而言，统计学分析包括两大基本方法：数字分析和图形分析。图形分析最适用于鉴定数据的表现形式，而数字分析更精确和客观。这两者互为补充且都必需。检验假设(hypothesis testing，常用 H 表示)是统计推断(inference)的重要方面。假设检验又称显著性检验(significance test)，是先对总体特征作出某种假设，然后从总体中抽样，计算相关统计量(statistic)，再根据检验结果作出拒绝或接受假设的判断。假设检验分参数检验(parametric test)和非参数检验(non-parametric test)两大类。10.3 节中将主要阐述常用的参数检验方法。

10.2.1 均数与方差

10.2.1.1 平均数

最常用于表示集中趋势的指标是算术平均数(arithmetic average)或简称平均数(mean)或均数。对于未分组资料，平均数是将所有数值相加再除以数值的个数。例如，现有以下 8 个观察值：

8　5　7　17　9　5　20　19　　(数组 A)

其平均数的计算方法为：

$$\text{平均数} = \frac{8+5+7+17+9+5+20+19}{8} = \frac{90}{8} = 11.25$$

平均数的通用计算公式中，通常须区分为总体平均数(μ)和样本平均数($\bar{X}$)。总体平均数是整个总体所有数值的平均数。样本平均数是总体抽样所得到样本的平均数。分别表示如下：

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \text{(公式 10-1)}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(公式 10-2)}$$

总体平均数和样本平均数概念的区分对于下文将要叙述的统计推断非常重要，其结论通常是根据所观察的样本来对总体进行估计。

10.2.1.2 样式

样式(mode)一词源于法语，表示时尚(fashion)的含义，在统计学中用来表示数据集(dataset)中最常见的数据。例如，表 10-2 或图 10-1 中，100 个基因表达值具有明显聚集在低表达水平(<100 单位)的特征。集中趋势的反映可用某个最常出现的数值来度量。当面对某些按区间分组的资料时，样式常用频数最高组的区间中点来表示。例如表 10-2 中，样式的表示是取 0~100 组的中点 50 来表示。最简便的样式计算方法是通过直方图来计算。

10.2.1.3 中位数

中位数(median)是将数据集由小到大排列后所得到的中间位置的数值，例如将上文平均数中的一组 8 个数据按由小到大次序排列，结果如下：

5 5 7 8 9 17 19 20 (数组 B)

其中，“8”和“9”位于排序的最中间位置，因此中位数表示为：

$$\text{中位数} = \frac{8+9}{2} = 8.5$$

中位数也可理解为小于数据集中 50% 数据但又大于数据集中另外 50% 数据的那个数值。

平均数、样式和中位数都是用来描述资料的集中趋势，它们之间也存在相互联系。对于一种分布对称的资料，这三者是相同的。如果是非对称分布资料，中位数总是介于平均数和样式之间。对于向右歪斜的资料，平均数最大，样式最小；对于向左歪斜的资料，则平均数最小，样式最大。

一个很自然的问题是，为什么需要使用中位数和样式来代替平均数？我们来看一个两张 cDNA 微阵列片所得到的数据标准化的例子。通常在比较两张片子的表达差异前，需要进行资料的调整。如果两张片子的数据呈线性相关，通常可用平均数或中位数来获得资料的集中趋势并进行各数据的标准化，但多数学者主张利用中位数进行标准化。这是因为一组非常大的 mRNA 表达强度的数据通常是向左歪斜的，仅少数属于高表达。如果所有基因的表达数据中，存在一些高表达或因灰尘污染导致的假高表达，对平均数的影响就较大。相反，中位数由于是按次序排列得到，所受到的噪音或极端值的影响就较小，或较不敏感。

10.2.1.4 方差

集中趋势的度量并不能反映数据分布的整个情况。具有相同均数的两组数据，其离散程度可完全不同。为了更好地反映资料的分布情况，需要描述资料的变异(variation)或离散(dispersion)。最简单的变异性度量方法是范围(range)，即资料的最小值和最大值之差。例如，数组 B 中，其范围为 20-5=15。范围越大，表明其变异性越大。范围的概念非常容易理解和计算，但没有考虑各数值或观察值之间的差异性。先看下面的一组资料：

5 9 10 11 11 12 12 20 (数组 C)

这组资料的范围数组 B 的范围完全相同，即都等于 15，但很明显的是，数组 C 的变异性要比数组 B 小。范围为 15 并不能反映两组数据的变异性，因此，需要新的指标来衡量。

最常用于衡量离散程度的指标是方差（variance，*Var*）及其平方根——标准差（standard deviation，*SD*）。其含义是指资料围绕均数的平均伸展。数组的变异越大，与均数之间的差值（简称离均差）越大。因此，变异的度量可用这些离均差的平均数来表示，即：

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n} \qquad \text{（公式 10－3）}$$

可以看出，公式 10－3 的结果总是为零，这不能用于比较两个数组的变异性。解决这一问题的方法是计算离均差的平方，再求总和，然后计算平均数，即按公式 10－4 表述。

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad \text{（公式 10－4）}$$

式 10－4 是表示总体方差的公式，如果是样本方差，其计算公式有所不同。即：

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad \text{（公式 10－5）}$$

值得注意的是，公式 10－5 中，如果分母采用 n 来计算，所得到的样本方差较总体方差偏低。校正方法是使分母采用（$n-1$）。此外，用公式 10－5 来估计总体方差时，总体大小 N 必须假定为无限大，才可能不出现偏差。实际应用中 N 至少应该为样本 n 的 50 倍以上。

由于方差采用了离均差的平方，如果对方差进行开平方来度量变异性，这种转换后所得到的结果就是标准差的计算方法。用数学表达式，即根据公式 10－4 和 10－5 分别得到总体标准差和样本标准差的计算公式：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} \qquad \text{（公式 10－6）}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad \text{（公式 10－7）}$$

标准差又被称为均方（mean square）的平方根。

标准差的计算举例中，我们继续来看数组 B 和数组 C。根据公式得到数组 B 和数组 C

的均数相同，即：

$$\bar{X} = 11.25$$

数组 B 的标准差根据公式 10－7，计算结果为：

$$s = \sqrt{\frac{(5-11.25)^2 + \cdots + (19-11.25)^2 + (20-11.25)^2}{8-1}} = \sqrt{\frac{281.5}{7}} = 6.34$$

而数组 C 的标准差的计算结果为：

$$s = \sqrt{\frac{(5-11.25)^2 + \cdots + (12-11.25)^2 + (20-11.25)^2}{8-1}} = \sqrt{\frac{123.5}{7}} = 4.2$$

通过以上计算，很显然，数组 B 的变异大于数组 C，尽管两者的均数相同。

10.2.1.5 变异系数

生物学研究中，经常遇到的一类问题是比较度量（scale）的差别。例如，需要比较两种不同的高通量 RNA 表达谱（如 cDNA 与高密度寡核苷酸微阵列）检测技术的可靠性，我们选定受试组中某种具有较高表达水平的基因 G1 来评估这两种技术。分别利用每一种技术在相同条件下重复测定 mRNA 表达量多次。用 A 技术测定时得到的均数和标准差分别为 150 和 25，而用 B 技术所得到的对应结果为 590 和 70。A 技术的标准差较 B 技术小，我们可能会得到技术 A 优于技术 B 的错误结论。

这种错误的产生是由于忽视了表达强度的不等分布，而仅依赖标准差的绝对值作出结论。为了避免这种错误的发生，引入变异系数（coefficiency of variance，*CV*）的概念，较标准差更好平均技术的优劣：

$$CV = \frac{s}{\bar{X}}$$

根据此公式重新计算上面例子中的结果，A 技术的 *CV* 为 17%，B 技术为 12%，说明 B 技术较 A 技术更可靠。

10.2.2 置信区间

上面 10.1.3 节中已经介绍了总体和抽样分布的概念，这里继续讨论与之相关的统计学概念。总体参数总是常数，而且只有一个这样的总体均数值 μ，但样本均数可有多个。从总体中抽取不同大小的不同样本时，可得到不同的样本均数 $\bar{X}$。这个样本均数属于随机变量，因为总体中每个成员具有相等的机会被抽取到样本中。例如，具有（1，2，3，4，5）这样 5 个样本的总体，其均数 μ 为 3.00，标准差 σ 为 1.58。若从该总体中随机抽取 3 个样本（不

置换),则可按下面的组合公式计算出 10 种不同的样本类型:

$$\text{样本类型总数} = \binom{5}{3} = \frac{5!}{3!\ (5-3)!} = \frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$$

其组合方式和相应的均数值列于表 10 - 3。从中可以看出,样本均数在 2.00 ~ 4.00 之间变化。

表 10 - 3　样本大小等于 3 时所有可能的样本和均数

总体	总体参数	样本	样本均数
1,2,3,4,5	均数 = 3.00	$S_1 = 1, 2, 3$	2.00
	标准差 = 1.58	$S_2 = 1, 2, 4$	2.33
		$S_3 = 1, 2, 5$	2.67
		$S_4 = 1, 3, 4$	2.67
		$S_5 = 1, 3, 5$	3.00
		$S_6 = 1, 4, 5$	3.33
		$S_7 = 2, 3, 4$	3.00
		$S_8 = 2, 3, 5$	3.33
		$S_9 = 2, 4, 5$	3.67
		$S_{10} = 3, 4, 5$	4.00
		平均样本均数	3.00
		样本均数标准差	0.61

抽样分布的均数就是所有样本均数($\bar{X}$)的平均数(即平均样本均数),用 $\mu_{\bar{x}}$ 表示。当从总体中包括所有可能的抽样类型时,其数值总是等于总体均数(μ),这一点可以从表 10 - 3 中看出,即总体均数和平均样本均数都等于 3.00。但样本均数的标准差(用 $\sigma_{\bar{x}}$ 表示)却不同于总体的标准差(σ),如表 10 - 3 中,这两个数值分别为 0.61 和 1.58。也就是说,抽样分布的方差小于其总体,这是抽样分布的重要特征,即中央极限定理(central limit theorem)。

如果变量 x 来自均数为 μ,标准差为 σ 的总体,其抽样分布的均数为 $\bar{X}$,随机样本数目为 n,则抽样分布的均数 $\mu_{\bar{x}}$ 和标准差 $\sigma_{\bar{x}}$ 分别表示为:

$$\mu_{\bar{x}} = \mu \text{ 和 } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

而且当样本大小足够大时,抽样分布符合正态分布。

抽样分布及中央极限定理的重要性在于抽样所得到的样本对总体进行推断,因为许多情况下,无法对总体的所有样本进行观察。这种推断包括两方面的内容:一是对总体参数的直接估计,这需要观察样本能大致代表真实的总体特征,如对总体均数的区间估计(*interval* estimation);二是观察样本是否支持对总体的先验假设(a priori hypothesis),即假设检验(hy-

pothesis testing)。这里先讨论区间估计,假设检验留待下一节专门讨论。

估计(estimation)是估计抽样所得到的信息对总体参数指定某个数值,包括点值估计和区间估计两种。点值估计是根据样本的统计资料而对总体参数的估计。例如通过样本均数($\overline{X}$)对总体均数(μ)的估计。区间估计有所不同,不是直接指定某个单一数值,而是围绕某一点值的大致区间的概率性描述,该区间包含总体参数在内。

由于样本大小不同,区间估计方法可有不同。当样本大小为30或更大时,可根据正态分布,利用样本均数来计算正态均数的置信区间(confidence interval, *CI*),区间的两个端值被称为置信限(confidence limit, *CL*)。其表达式为:

$$\mu = \overline{X} \pm z\sigma_{\overline{x}} \qquad (当\ \sigma\ 已知时)$$

$$\mu = \overline{X} \pm zs_{\overline{x}} \qquad (当\ \sigma\ 未知时)$$

其中,$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$和$s_{\overline{x}} = \frac{s}{\sqrt{n}}$。

z的数值可根据标准正态分布表查找。例如,要估计95%的置信区间时,$z = 1.96$,即正态分布中95%的观察值落在±1.96个标准差范围之内。其置信区间表示为:

$$(s_x - 1.96_{s_x}) < \mu < (s_{\overline{x}} + 1.96_{s_x}) \qquad (公式\ 10-8)$$

当样本大小小于30时,如果抽样分布仍然符合正态分布,而且总体的标准差($\delta_{\overline{X}}$)已知时,仍按上面公式10-8的正态分布方法进行估计。否则,须用t分布代替正态分布用于对总体均数(μ)的置信区间估计,其表达式为:

$$\mu = \overline{X} \pm ts_{\overline{x}},其中,s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

其中t值由自由度(degree of freedom, df)为$(n-1)$时的t值表中得到。

10.2.3 假设检验

上节中我们讨论了如何利用样本的统计资料帮助我们对总体参数进行估计的概念。接下来需要讨论的是利用假设检验进行统计推断的问题,即利用对样本的统计分析来支持某个关于总体参数的先验假设。通过观察样本来测试这种假设是否为真的可能性。从概率论角度分析,如果这种可能性不能维持,则另一种相对的假设被认为是真。这两类假设分别称为无效假设(null hypothesis, H_0)和备择假设(alternative hypothesis, H_1)。无效假设是指对总体参数假定为真,除非其被证实为假;而备择假设是当无效假设不成立时对总体参数假定为真。根据这种定义,假设检验最常用的方法是建立一套关于参数真值的相互排斥的假设,而样本统计就是用来支持这两种假设之一。

以下通过一个基因表达的例子来说明假设检验。在一个基因表达实验中,基因表达的

变化通常分为以下几种情形：

(1)在实验条件下，基因表达未发生变化，$\bar{X}=\mu$；

(2)在实验条件下，基因表达上调，$\bar{X}>\mu$；

(3)在实验条件下，基因表达下调，$\bar{X}<\mu$。

假设基因 G_1 的表达与肿瘤有关，利用一种叫做 X_1 的化合物处理后，分别测定 5 次。基因表达的平均数为 121。已知未进行 X_1 处理时基因表达水平的均数为 110，标准差为 12。那么，该化合物处理后，G_1 的表达是否上调？

这个例子中，我们注意到经化合物 X_1 处理后基因 G_1 的表达较未处理时高。问题是这种变化是否具有生物学意义，还是因为随机因素的影响？这种情况下，有以下两种假设：①无效假设(H_0)：处理样品的平均基因表达水平小于或等于未处理样品的基因表达水平，即 $H_0:\bar{X}\leqslant\mu$；②研究假设(H_1)：处理样品的平均基因表达水平高于未处理样品的基因表达水平，即 $H_1:\bar{X}>\mu$。

这两个假设包含所有的可能性。数值预期出现在分布的右侧(单侧检验)。根据样本分布定理，样本均数呈正态分布，因此可用正态分布进行假设检验。已知：

$$n=5,\bar{X}=121,\mu=110,\sigma=12$$

对于 $\bar{X}$ 检验统计量(test statistic)z 的计算方法如下：

$$z=\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{(公式 10-9)}$$

则

$$z=\frac{121-110}{\frac{12}{\sqrt{5}}}=\frac{11}{5.37}=2.05$$

根据统计学书中的标准正态分布表，可以看出 z 值的变化显著性水平为：$\alpha=0.01$ 和 $\alpha=0.05$ 时，z 值分别为 2.33 和 1.65。当选择显著性水平为 0.01 时，检验统计量 $z=2.05$，低于其界值(critical value) 2.33。因此，认为样本均数距离总体均数很近，经 X_1 化合物处理后，G_1 基因的表达没有上调。但是，如果选择显著性水平为 0.05，检验统计量 $z=2.05$，则大于界值 1.65。这种情况下，认为样本均数距离总体均数足够远，即经 X_1 化合物处理后，G_1 基因的表达出现上调。

通常，显著性水平是在计算检验统计量之前预先设置好的，假设检验遵循以下步骤：①提出无效假设(H_0)和研究假设(H_1)，这两种假设必须是相互排斥并包含所有可能性；②选择合适的分布进行分析；③选择显著性水平 α；④根据资料计算检验统计量；⑤比较某一给定显著性水平 α 下的界值与检验统计量的大小，作出是否接受无效假设的决定。

值得注意的是，以上的假设检验举例中，在选择不同的显著性水平时，同一组资料，可能得到上调或无变化的结论。这有必要了解以下有关假设检验错误的概念。

10.2.4 Ⅰ类和Ⅱ类错误

许多教科书中用来说明假设检验错误的最著名的例子是法庭中判断一个人是否有罪的

例子。这里,我们沿用上节假设检验中的例子来解释假设检验的错误。先有必要提醒读者回顾一下该例中的两个假设:无效假设和研究假设。

基因表达的实际情况中,总是存在两种情形,一种是基因表达量不增加,另一种是基因表达量增加。在前一种情形下,即 H_0 为真时,有两种假设检验的结果:①基因事实上是未上调,但被认为是上调,即 H_0 为真但 H_0 被拒绝;②基因事实上是未上调,也被认为未上调,即 H_0 为真但 H_0 未被拒绝。

以上第一种情况,已得出基因上调的错误结论,但事实上该基因未上调,即出现假阳性(false positive)。统计学上称这类错误为Ⅰ类错误(type Ⅰ error)。第二种情况,结论反映了真实情况,即真阴性(true negative)。Ⅰ类错误中,假设检验的显著性水平 α 反映了事实上为真而拒绝无效假设的概率,这决定了被拒绝区域的大小,这是为什么上一节的例子出现两种不同结论的原因。

同样,当实际情况为基因表达增加时,也会出现两种检验假设的结果:①基因事实上是上调,也被认为如此,即 H_0 为假而且 H_0 被拒绝;②基因事实上是上调,但被认为未上调,即 H_0 为假但 H_0 未被拒绝。

前一种情况结论是正确的,即真阳性(true positive)。后一种情况,基因实际上是增加的,我们却得到一个错误的判断认为未上调,属于假阴性(false negative),因此失去获得一种潜在的抗癌药的机会。统计学上称这类错误为Ⅱ类错误(type Ⅱ error)或 β 错误。真阳性率($1-\beta$)又被称为检验效能(power of test)。这两类错误总结见表 10-4。

表 10-4 假设检验的Ⅰ、Ⅱ类错误

统计检验结论	真实状态	
	H_0 为真	H_0 为假
H_0 为真(H_0 被接受)	结论正确(真阴性,$1-\alpha$)	Ⅱ类错误(假阴性,β)
H_0 为假(H_0 被拒绝)	Ⅰ类错误(假阳性,α)	结论正确(真阳性,$1-\beta$)

这两类错误相互依赖,对于固定样本大小的假设检验,不可能同时降低 α 和 β。降低 α 会增加 β,反过来亦然。虽然很容易判断Ⅰ类错误的大小,因为其值等于显著性水平 α,但计算Ⅱ类错误的概率并不容易。除与显著性水平有关外,Ⅱ类错误还与其他因素如总体标准差(σ)、样本大小(决定抽样分布的标准误)等有关。通常,提高Ⅰ类错误率(α)、增大假定与真实值之间的差异或增加样本量可降低Ⅱ类错误率。

10.3 常用统计检验方法

统计检验方法有多种。通常分为参数检验和非参数检验两大类。参数检验依赖于分布,而非参数检验可以在数据分布不明确时采用。这里仅介绍参数检验中最常用的 t 检验(student t test)、方差分析(analysis of variance, ANOVA)。卡方检验(chi-square test)也是一种常用的参数检验方法,将在 10.4 节"关联分析"中介绍。

10.3.1 t 检验

前面 10.2.3 节的假设检验中，是比较样本均数与总体均数的差异，又称 Z 检验。当样本分布符合 t 分布（小样本）时，所涉及的 Z 检验方法即为 t 检验的一种。t 检验还常用于比较两个总体均数是否存在差异。

对于两个总体的几种假设检验分如下几种不同情况：①检验两个大样本和相对独立的总体均数的差异；②检验两个小样本和相对独立且标准差相同的总体均数的差异；③检验两个小样本和相对独立但标准差不同的总体均数的差异；④检验两个配对样本的总体均数的差异。

这里所谓的小样本，通常是指样本量小于 30。第 1 种情况仍可用上述 Z 检验方法或 U 检验方法。这里主要讨论第②、③两种情形下的 t 检验方法。

t 检验中，必须注意区分配对或不配对资料，方差相同或方差不齐，单侧检验或双侧检验。单侧检验只限于具有先验知识的情况下，否则采用双侧检验。

先看下面的例子：

假如某个与肿瘤有关的基因 G_1，分别在处理组（加药物 X_1）和对照组（不加药物 X_1）中进行测定。在对照组中所得结果为 45，32，62，34，59，55，39，33，在处理组中所得结果为 34，35，40，22，31，40，37，32，39，42，那么基因的表达在两组间是否存在差异？

这个例子中，我们并不知道两个样本的方差是否相同（以 s_c^2 表示对照组方差，s_t^2 表示处理组方差）。即使这两个样本来自同一总体，其方差很可能不同。这样，我们需要建立一种方法进行检验和推断。为了比较方差的齐性，通常比较两种方差的比值 $\frac{s_c^2}{s_t^2}$ 或 $\frac{s_t^2}{s_c^2}$。这个比值符合 F 分布。这个例子中，可得到以下统计结果：

$n_c = 8$　　$\bar{X}_c = 45.87$　　$s_c^2 = 150.70$　　$df_c = 7$

$n_t = 10$　　$\bar{X}_t = 35.2,$　　$s_t^2 = 34.84$　　$df_t = 9$

比较两组样本方差的假设检验为：

$$H_0: \sigma_c = \sigma_t$$

$$H_a: \sigma_c \neq \sigma_t$$

显著性水平被设定为 0.05 或 0.02（为描述方便）。根据以上假设，属于双侧检验，必须考虑 F 分布的两端，为了简化其过程，我们只考虑右侧 F 分布的情形，即 $\frac{s_c^2}{s_t^2} \geq 1$。根据 F 分布表，得到：

$$F_{(/2\,(df,\,df)} = F_{0.025\,(7,\,9)} = 4.20$$

$$F_{(/2\,(df,\,df)} = F_{0.01\,(7,\,9)} = 5.61$$

如果方差比值大于 F 值，H_0被拒绝。方差比值结果计算如下：

$$\frac{s_c^2}{s_t^2} = \frac{150.70}{34.84} = 4.35$$

当显著性水平设定为0.05时，H_0被拒绝(4.35 > 4.20)，表明两者的方差不齐。但是，当显著性水平设定为0.02时，H_0被接受(4.35 < 5.61)，表明两者方差相同。这两种假设检验结果可导致使用不同的均数比较方法。对于方差相同或不齐时假设检验的一般步骤如下。

(1)无效假设和备择假设：

H_0： $\mu_c = \mu_t$

H_a： $\mu_c \neq \mu_t$

(2)对于小样本(样本量小于30)，符合 t 分布，采用 t 检验；

(3)显著性水平设定为 $\alpha = 0.05$，备择假设表明属于双侧检验，每侧的显著性水平都等于：

$$\alpha/2 = 0.025$$

(4)t 检验统计量为：

$$t = \frac{(\bar{x}_c - \bar{x}_t) - (\mu_c - \mu_t)}{s_{\bar{x}_c - \bar{x}_t}}$$

其中 $s_{\bar{x}_c - \bar{x}_t}$表示两样本均数的估计标准差，根据无效假设 $\mu_c = \mu_t$，因此 t 检验统计量的计算公式为：

$$t = \frac{\bar{x}_c - \bar{x}_t}{s_{\bar{x}_c - \bar{x}_t}} \qquad \text{(公式 10-10)}$$

10.3.1.1 方差相同时两独立的小样本的假设检验

上例中，如果考虑两样本方差相同，式10-10中的参数计算如下：

$$s_{\bar{x}_c-\bar{x}_t}=s_p\sqrt{\frac{1}{n}+\frac{1}{n}} \qquad \text{(公式 10-11)}$$

$$s_p=\sqrt{\frac{(n_c-1)s_c^2+(n_t-1)s_t^2}{n_c+n_t-2}} \qquad \text{(公式 10-12)}$$

其中 S_p表示两样本的合并标准差。将有关数值分别代入公式 10-10、10-11 和 10-12 得到的计算结果为：

$$s_p=\sqrt{\frac{(8-1)\times 150.70+(10-1)\times 34.84}{8+10-2}}=9.25$$

$$s_{\bar{x}_c-\bar{x}_t}=9.25\times\sqrt{\frac{1}{8}+\frac{1}{10}}=9.25\times 0.47=4.35$$

$$t=\frac{44.88-35.20}{4.35}=2.23$$

$$df=n_c+n_t-2=8+10-2=16$$

当 df=16 和显著性水平为 0.05 时，双侧 t 分布的界值为 -2.12 和 2.12。由于 t 统计值大于 2.12，无效假设被拒绝。这里的结论是处理组和不处理组间 G_1 基因的表达差异具有显著性意义。

10.3.1.2 方差相同时两独立的小样本的假设检验

若考虑上例中两样本方差不齐，有关 t 统计量计算的相关参数为：

$$s_{\bar{x}_c-\bar{x}_t}=\sqrt{\frac{s_c^2}{n_c}+\frac{s_t^2}{n_t}} \qquad \text{(公式 10-13)}$$

$$s_{\bar{x}_c-\bar{x}_t}=\sqrt{\frac{150.70}{8}+\frac{34.84}{10}}=4.72$$

$$t=\frac{44.88-35.20}{4.72}=2.05$$

$$df = \frac{\left(\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}\right)^2}{\frac{\left(\frac{s_c^2}{n_c}\right)^2}{n_c - 1} + \frac{\left(\frac{s_t^2}{n_t}\right)^2}{n_c - 1}} \quad \text{（公式 10-14）}$$

$$df = \frac{\left(\frac{150.7}{8} + \frac{34.84}{10}\right)^2}{\frac{\left(\frac{150.7}{8}\right)^2}{8-1} + \frac{\left(\frac{34.84}{10}\right)^2}{10-1}} = 9.5 \approx 10$$

当 df = 10 和显著性水平为 0.05 时，双侧 t 分布的界值为 - 2.23 和 2.23。由于 t 统计值小于 2.12，无效假设被接受。这里的结论是处理组和不处理组间 G_1 基因的表达差异不具有显著性意义。

以上结果表明，假设检验的每一环节都对推断非常重要。

10.3.2 方差分析

t 检验主要用来对小样本两均数的假设检验。实际应用中常涉及多组资料的比较。这时尽管可用 t 检验进行两两比较，但相对烦琐。这时适合选用方差分析（ANOVA）。

方差分析是建立在 F 分布的基础上，其假设是：①资料符合正态分布；②各组间的方差必须等同或非常接近；③各观察值应该相互独立。

ANOVA 分析的无效假设是多个总体间的均数没有差别，其对应的备择假设为不是所有总体的均数都相同。当然，ANOVA 也可用于比较两总体均数的差异性，但这时使用 t 检验的效率更高。ANOVA 的检验假设为：

H_0 = 所有总体均数都相同（$H_0: \mu_1 = \mu_2 = \cdots = \mu_n$）。

H_1 = 不是所有总体均数都相同。

备择假设中，可分为以下几种不同情况：①所有均数间互不相同；②其中一些但不是所有的均数存在差异；③至少一对均数存在差异。

ANOVA 分单因素（one - way）和双因素（two - way）两种，它们的实验设计和资料分析方法有所不同，但两者的核心都是将方差进行拆分。这里仅介绍单因素 ANOVA。先看下面的例子：

假设与肿瘤相关的基因 G_1，分别用 cDNA 微阵列测定其在对照组（未经处理）、处理组 1（经化合物 X_1 处理）和处理组 2（经化合物 X_2 处理）中的表达共 10 次，所得结果见表 10-5。需要知道，基因 G_1 在这 3 种不同条件下，其表达是否存在差异？

表 10－5　基因 G_i 在不同处理条件下的表达量和 ANOVA 分析表

微阵列	对照组	处理组 X_1	处理组 X_2
1	45	34	49
2	39	35	51
3	49	40	30
4	34	22	61
5	55	31	57
6	59	40	39
7	39	37	33
8	30	32	42
9	45	39	40
10	36	42	58
T_i	$T_c=431$	$T_{x1}=352$	$T_{x2}=460$
$\overline{X}_i$	$\overline{X}_c=43.1$	$\overline{X}_{x1}=35.2$	$\overline{X}_{x2}=46.0$
s_i^2	$s_c^2=86.1$	$s_{x1}^2=34.8$	$s_{x2}^2=116.6$
n_i	$n_c=10$	$n_{x1}=10$	$n_{x2}=10$

其中，T_i表示各样品数值总和，n_i表示样本大小。

ANOVA 分析步骤如下：

(1)作出无效假设和备择假设：

H_0 = 所有 3 个均数都相同(H_0:$\mu_1=\mu_2=\mu_3$)；

H_1 = 3 个均数并不都相同。

(2)确定显著性水平为 0.05。

(3)计算 F 统计量：先计算组间平方和(SSB)和组内平方和(SSW)，然后计算组间方差(MSB)和组内方差(MSW)。

$$SSB=\left(\frac{T_1^2}{n_1}+\frac{T_2^2}{n_2}+\cdots+\frac{T_n^2}{n_n}\right)-\frac{(\Sigma x)^2}{n} \qquad (公式\ 10-15)$$

$$SSB=\left[\frac{(431)^2}{10}+\frac{(352)^2}{10}+\frac{(460)^2}{10}\right]-\frac{(431+352+460)^2}{30}=624.87$$

$$SSW=\Sigma x^2-\left(\frac{T_1^2}{n_1}+\frac{T_2^2}{n_2}+\cdots+\frac{T_n^2}{n_n}\right) \qquad (公式\ 10-16)$$

$$SSW=(45)^2+(39)^2+\cdots+(58)^2-\left[\frac{(431)^2}{10}+\frac{(352)^2}{10}+\frac{(460)^2}{10}\right]=2\ 138.5$$

$$MSB=\frac{SSB}{k-1} \qquad (公式\ 10-17)$$

其中 $k-1$ 是组间方差的自由度，即 $3-1=2$. 因此：

$$\mathrm{MSB}=\frac{624.84}{3-1}=312.43$$

$$\mathrm{MSW}=\frac{\mathrm{SSW}}{n-k} \qquad \text{（公式 10-18）}$$

其中 $n-k$ 为组内方差的自由度，即 $30-3=27$，因此：

$$\mathrm{MSW}=\frac{2138.5}{30-3}=79.2$$

F 统计量的计算结果为：

$$F=\frac{\mathrm{MSB}}{\mathrm{MSW}}=\frac{312.43}{79.2}=3.94$$

许多统计软件可非常便利地直接给出 ANOVA 的分析结果。如此例分析中，用微软的 Excel 软件得到的 ANOVA 统计结果见表 10-6。

表 10-6　Excel 输出的 ANOVA 分析结果

变异来源	SS	df	MS	F	P-value	F 界值
组间变异	624.866 7	2	312.433 3	3.944 681	0.031 409	3.354 131
组内变异	2 138.5	27	79.2 037			
总变异	2 763.367	29				

(4)作出统计推断或结论。由于计算得到的 $F=3.94$，大于临界值 3.35，因此无效假设被拒绝，认为基因表达结果中，至少两组的表达差异具有显著性意义。

（胡自华　著　李　巍　译）

10.4　关联分析

关联(association)是指两个事件非随机地共同发生(正关联)或非随机地相互排斥(负关联)。关联分析是统计遗传学和药物遗传学中的常用分析方法，主要用于病例对照资料的分析。其中最常用的统计方法是相对危险度(relative risk，RR，或称相对危险率)和几率比值(odds ration，OR)的计算。在讨论这些计算方法前，须明确两个统计学基本概念：患病率(prevalence，P)与发病率(incidence，I)。患病率是指某一观察时点总人群中存在的患者总人数。发病率是指某一观察窗口时间内，危险人群中新发生的病例数。患病率通常由回顾性研究(retrospective study)得到，发病率通常由前瞻性研究(prospective study)得到。RR 的计算是采用发病率，OR 的计算常用患病率，也可用发病率。

10.4.1　相对危险度[3,4]

相对危险度又称相对发病率(relative incidence)，是指具有某种危险因素(exposed risk)

人群中某种疾病的发病率与不具有这种危险因素人群中该疾病发病率的比值，即具有危险因素者某疾病的发生危险率是无危险因素者的倍数。所谓危险因素通常指要观察的发病风险因素（如吸烟、高血压）、基因型（如 HLA 分型）或药物遗传性状（如 SNP）等。RR 分析主要用于病例对照资料。首先，必须设定病例和对照两个组别，再分别统计每组中具有或不具有某种危险因素的人数，列出表 10－7 的四格表（2×2 表）。

表 10－7　相对危险度分析表

危险因素 RISK	病例数 CASE	对照数 CONTROL	合计 TOTAL
有（YES）	a	b	a＋b＝E
无（NO）	c	d	c＋d＝F
合计（TOTAL）	a＋c＝G	b＋d＝H	a＋b＋c＋d＝N

根据表 10－7，RR 的计算公式为：

$$RR = (a/E)/(c/F) = aF/cE \qquad (公式\ 10-19)$$

计算结果的分析中：

当 RR＝1，无关联性；

当 RR＞1，正关联性，而且 RR 越大，正关联程度越高；

当 RR＜1，负关联性，而且 RR 越小，负关联程度越高。

在作统计推断前，须对所得到的 RR 值进行显著性检验。但结果分析时，更重要的是计算 RR 的置信限（confidence limit，*CL*），常用 Katz 法。

$$Var(\ln RR) = [(b/a)/(b+a)] + [(d/c)/(c+d)] = [SE(\ln RR)]^2 \qquad (公式\ 10-20)$$

$$95\%\ CL = \ln RR \pm 1.96SE(\ln RR) \qquad (公式\ 10-21)$$

根据 RR 的置信限，判断其关联性是否具有统计学意义。

10.4.2　几率比值[4]

几率比值的计算中，数据资料与 RR 相似，见四格表 10－7。OR 是指具有危险因素人群中患者与非患者之比相当于不具有危险因素人群中患者与非患者之比的倍数，即患者与非患者的几率在危险和非危险人群中的比值。OR 的计算公式为

$$OR = (a/b)/(c/d) = ad/bc \qquad (公式\ 10-22)$$

值得注意的是，如果将式 10－22 作适当变换，可以看出：

$$(a/c)/(b/d) = (d/c)/(b/a) = (d/b)/(c/a) = ad/bc = OR \quad (式\ 10-23)$$

式 10－23 代表不同的配对比较，得到的 OR 相同。如病例组中有危险因素的患者与无危险因素的患者之比(a/c)相当于对照组中有危险因素的非患者与无危险因素的非患者之比(b/d)的倍数，也同样等于上述 OR 值。

一般地，如果某种疾病属罕见病，即患病率 P 通常小于 0.05 时，常用 OR 代替 RR。

OR 计算结果的分析与 RR 相似：

当 OR＝1，无关联性；

当 OR＞1，正关联性，而且 OR 越大，正关联程度越高；

当 OR＜1，负关联性，而且 OR 越小，负关联程度越高。

在作统计推断前，须对所得到的 OR 值进行显著性检验。同样，结果分析时，须计算 OR 的置信限。常用 Woolf 法。

$$Var(\ln OR) = (1/a) + (1/b) + (1/c) + (1/d) = [SE(\ln OR)]^2 \quad (公式\ 10-24)$$

$$95\%\ CL = \ln OR(1.96SE(\ln OR) \quad (公式\ 10-25)$$

例如，在有关冠心病与血清胆固醇水平的研究中，得到表 10－8 的结果。

表 10－8　血清胆固醇水平与冠心病的发生

血清胆固醇 (mg/dL)	冠心病患者数	对照数	合计
≥ 250	a＝10	b＝125	a＋b＝135
＜ 250	c＝21	d＝449	c＋d＝470
合计	a＋c＝31	b＋d＝574	a＋b＋c＋d＝605

根据公式 10－19，得到 RR＝1.66；根据公式 10－22，得到 OR＝1.71。RR 和 OR 都大于 1.0，似乎认为血清胆固醇水平高于 2.5 g/L 时，与冠心病的发生呈正关联性。进一步计算 OR 的 CL。根据公式 10－24，求得 $SE(\ln OR)$＝0.397 3，$\ln OR = \ln 1.71$＝0.536 5，因此按公式 10－23，95% CL 为 0.79～3.73。同样按公式 10－20 和 10－21，可求得 RR 的 95% CL 为 0.80～3.43。因 OR 和 RR 的置信限都包含 1.0，认为这种关联性在统计学上无意义。以上的统计结果可用 SAS 进行验证。

10.4.3　归因危险度[4]

归因危险度(attributable risk，AR)是指总的相对危险度中，要求分析由于危险因素所引起的危险度是多大，以及该危险因素对整个人群的危险性又是多大。

为了叙述方便,先结合表 10－7 给出下列符号所代表的含义:

I_r = 危险组的发病率(a/E);

I_0 = 非危险组的发病率(c/F);

RR = 相对危险度(I_r/I_0);

I_t = 总人群中的发病率(G/N);

P = 危险组人群占总人群的比例(E/N);

I_{ex} = 危险组中因危险因素所引起的发病率;

n_{ex} = 由于危险因素所引起的病例数;

$AR_{EC}\%$ = 危险因素对危险组的归因危险度百分数;

$AR_{total}\%$ = 危险因素对总人群的归因危险度百分数。

根据以下公式分别计算:

$$I_{ex} = I_r - I_0 \qquad \text{(公式 10－26)}$$

$$n_{ex} = a \times (RR - 1)/RR, a = NPI_r \qquad \text{(公式 10－27)}$$

$$AR_{EC}\% = (I_{ex}/I_r) \times 100 = [(RR - 1)/RR] \times 100 \qquad \text{(公式 10－28)}$$

$$AR_{total}\% = [(I_t - I_0)/I_t] \times 100 \qquad \text{(公式 10－29a)}$$

$$= [P(RR - 1) \times 100/[1 + P(RR - 1)] \qquad \text{(公式 10－29b)}$$

公式 10－29a 用于队列研究(cohort study),公式 10－29b 用于队列研究或病例对照研究。这些统计的假设是没有其他混杂因素的干扰。以表 10－9 的 18 年期收缩压与冠心病关联性队列研究结果为例,说明归因危险度的计算方法。

表 10－9　收缩压(SBP)与冠心病(CHD)关联性调查结果(18 年期)

SBP (mmHg)	CHD 患者		危险人群数	发病率	RR
	是	否			
≥165	95	201	296	0.321	1.98
<165	173	894	1 067	0.162	1.00
合计	268	1 095	1 363	0.197	/

根据上表结果,可以得到以下分析数据:

(1) $I_{ex} = I_r - I_0 = 0.321 - 0.162 = 0.159$,即高 SBP 人群(≥165 mmHg)中因 SBP 高引起的 CHD 发病率为 0.159。

(2) $n_{ex} = a \times (RR - 1)/RR = 95 \times 0.98/1.98 = 47$,即高 SBP 人群中因 SBP 高导致的

CHD 患病人数为 47 人。

(3) $AR_{EC}\% = (I_{ex}/I_r) \times 100\% = 0.159 \times 100\%/0.321 = 49.5\%$，表明高 SBP 人群中 SBP 高对 CHD 的贡献率为 49.5%。

(4) $AR_{total}\% = [(I_t - I_0)/I_t] \times 100\% = (0.197 - 0.162) \times 100\%/0.197 = 17.8\%$，表明由于 SBP 高(165 mmHg)引起的冠心病占人群中所有冠心病的 17.8%。

如果危险因素分多个不同水平(strata)，则按以下公式计算总人群归因危险度(AR_T)和各不同水平的归因危险度(AR_i)。

$$AR_T = \Sigma[(P_i \times RR_i - P_i)/\Sigma(P_i \times RR_i)] = \Sigma AR_i = (I_T - I_0)/I_T, I_T = \Sigma(P_i \times I_i) \quad (公式\ 10-30)$$

其中，I_i、P_i和 RR_i分别表示第"i"水平的发病率、人群比例与相对危险度。

$$AR_i = (P_i \times RR_i - P_i)/\Sigma(P_i \times RR_i) \quad (公式\ 10-31)$$

仍以表 10-9 的研究为例，说明其计算方法。这里 SBP < 165 的组别被分成 140~164 和 <140 两个组，即 SBP 共分 3 个不同水平，见表 10-10 结果。

表 10-10 收缩压(SBP)与冠心病(CHD)关联性调查与统计结果(18 年期)

SBP (mmHg)	CHD 患者		合计	I_i	P_i	RR_i	AR_i
	是	否					
≥165	95	201	296	0.321	0.217	2.50	0.214
140~160	100	399	499	0.200	0.366	1.56	0.133
<140	73	495	568	0.129	0.417	1.00	0.000
合计	268	1 095	1 363	0.197	1.000	/	0.347

根据表 10-10 的结果及公式 10-30，可得到以下结果：

(1)总人群归因危险度 $AR_T = (I_T - I_0)/I_T = (0.197 - 0.129)/0.197 = 0.345$，表明由于 SBP 高(≥140 mmHg)引起的冠心病占人群中所有冠心病的 34.5%。

(2)不同 SBP 水平的归因危险度的计算(按公式 10-31)：

$$\Sigma P_i \times RR_i = 0.219 \times 2.50 + 0.366 \times 1.56 + 0.417 \times 1.00 = 1.535$$

$$AR_{\geq 165} = (0.219 \times 2.50 - 0.219)/1.535 = 0.214$$

$$AR_{140-164} = (0.366 \times 1.56 - 0.366)/1.535 = 0.133$$

$$AR_{<140} = (0.417 \times 1.00 - 0.417)/1.535 = 0.000$$

这里可以根据(2)验证(1)的结果，即：

$$AR_T = \Sigma AR_i = 0.214 + 0.133 + 0.000 = 0.347$$

10.4.4 率的标准化[4]

当人群结构出现差异时，如年龄、性别的不同，率的比较会出现偏差。因此在率的比较

前，必须先对率进行标准化(standardization)。最常用的率的校正方法或标准化方法是对年龄的校正，分直接校正和间接校正两种方法。

10.4.4.1 率的直接校正法

直接校正法是在已知标准人群各年龄段人群数(n_{is})和待校正人群各年龄段率(p_i)的基础上进行的。这样，两个或多个率都以标准人群为参照，校正率就排除了年龄结构的不同，可以较客观地反映率的差别与否。其校正公式为：

$$\text{校正率} = \Sigma\ (p_i\ n_{is}) / \Sigma\ n_{is} \qquad \text{(公式 10-32)}$$

10.4.4.2 率的间接校正法

间接校正法是在已知待校正人群各年龄段率(p_{ia})与人群数(n_{ia})，以及标准人群中各年龄段率(p_{is})和人群数(n_{is})的基础上的校正方法。主要解决的问题是：在与标准人群相同年龄段发病率(morbidity)或死亡率(mortality)的情况下，被研究人群中各年龄段期望的发病或死亡人数。间接校正中需要先计算标准化发病率或死亡率(简称 SMR)：

$$\begin{aligned}\text{SMR} &= \text{观察人数总和/期望人数总和} \\ &= \Sigma\ (p_{ia}n_{ia}) / \Sigma\ (p_{is}n_{ia}) \end{aligned} \qquad \text{(公式 10-33)}$$

$$\text{校正率} = \text{SMR} \times \Sigma\ (p_{is}n_{is}) / \Sigma\ n_{is} \qquad \text{(公式 10-34)}$$

公式 10-33 的计算结果中，当 SMR > 1 时，表示相对标准人群而言，所研究人群的危险度更高；SMR < 1 时，表明所研究人群危险度较标准人群低；SMR = 1 时，表示两者无区别。实际应用中，尚须计算 SMR 的置信限，其方差为(SMR/E)，E 为期望人数总和。

10.4.5 几率比值的校正[2,4]

前面 10.4.2 节有关 OR 的计算是假定无混杂因素(confounder)或相互作用(interaction)的情况。实际应用中，OR 常受年龄、性别等的影响，可导致偏差，必须进行校正，才能较客观地反映关联性大小。这里所用的校正方法是指不用多因素分析模型进行的校正，常用 Mantel-Haenzel 和 Woolf 两种方法。实际应用中，对于较复杂的样本资料，常用下文将叙及的多元回归模型进行校正，以排除混杂因素和相互作用的影响。

10.4.5.1 Mantel-Haenzel 校正法

Mantel-Haenzel 法(简称 M-H 法)对 OR 的校正是根据可能的混杂因素(如年龄)将资料拆分为不同年龄段的 2×2 资料(strata)，再根据这些不同分层资料(符号意义参见表 10-7)，利用公式 10-35 计算总的 OR(OR_{MH})，然后按公式 10-36 进行显著性检验(χ^2_{MH}的自由度为 1)，再按公式 10-37 计算该 OR 的置信限。利用 SAS 软件的 CMH 选项时，可自动给出卡方检验结果和 M-H 法校正 OR 的置信限(见 OR 结果中 Mantel-Haenzel 一栏)，详见 10.6.2 节。

$$OR_{MH} = \Sigma(a_i d_i / N_i) / (\Sigma b_i c_i / N_i) \quad \text{（公式 10-35）}$$

$$\chi^2_{MH} = (\Sigma a_i - \Sigma m_i)^2 / \Sigma v_i \quad \text{（公式 10-36）}$$

其中

$$m_i = E_i \times G_i / N_i$$

$$v_i = (E_i \times F_i \times G_i \times H_i) / [N^{2i} \times (N_i - 1)]$$

$$95\%CL = \ln OR_{MH} \pm 1.96[\ln OR_{MH} / (\chi^2_{MH})^{1/2}] \quad \text{（公式 10-37）}$$

10.4.5.2 Woolf 校正法

Woolf 法对 OR 的校正较适用于两个以上混杂因素（如年龄+性别）的资料，并可测试不同混杂因素间是否存在相互作用（公式 10-40），但不能直接用于单元格数据为“0”的资料。这时，可用数值“0.5”来替代，或选用 M-H 法。OR_{Woolf} 的计算方法见公式 10-38，其置信限的计算用公式 10-39，利用 SAS 软件的 CMH 选项时，也可同时给出 Woolf 法的校正 OR 置信限（见 OR 结果中 Logit 一栏）。

$$\ln OR_{Woolf} = \Sigma(w_i \times \ln OR_i) / \Sigma w_i \quad \text{（公式 10-38）}$$

其中

$$w_i = (1/a_i + 1/b_i + 1/c_i + 1/d_i)^{-1} = 1/Var(\ln OR_i)$$

$$95\%CL = \ln OR_{Woolf} \pm 1.96 \times [1/(w_i)^{1/2}] \quad \text{（公式 10-39）}$$

$$\chi^2_{k-1} = (\ln OR_i - \ln OR_{Woolf})^2 / Var(\ln OR_i) \quad \text{（公式 10-40）}$$

自由度中，k 表示 2×2 表格数。

例如，表 10-10 中，如果将 SBP 分为≥140 和<140 两组，根据公式 10-22 和公式10-25 可分别求得校正前的 OR 值=(195×495)/(73×600)=2.20，95%CL 为 1.64~2.95。

现将该组资料按年龄和性别细分为表 10-11 的结果，即被拆分为 4 个 2×2 表格。

表 10-11 不同性别和年龄组收缩压（SBP）与冠心病关联性调查结果

性别	年龄	SBP(mmHg)	患者数	非患者数	合计
男性	≥55	≥140	50	85	135
		<140	20	78	98
	<55	≥140	58	154	212
		<140	36	162	198
女性	≥55	≥140	46	143	189
		<140	7	64	71
	<55	≥140	41	218	259
		<140	10	191	201
合计			268	1 095	1 363

根据上表资料，可将计算结果列于表 10－12。

表 10－12　Woolf 法校正 OR 的统计参数表

组别	OR_i	$\ln OR_i$	$Var(\ln OR_i)$	$1/w_i$
男性≥55	2.294	0.830	0.095	10.572
女性<55	1.695	0.528	0.058	17.335
男性≥55	2.941	1.079	0.187	5.341
女性<55	3.592	1.279	0.134	7.451

将表 10－12 的计算结果代入式 10－38，得到 $\ln OR_{Woolf}=0.816$，即 $OR_{Woolf}=2.26$。校正后的 OR 值与校正前的 OR 值(2.20)非常接近。再按公式 10－39，计算出 OR_{Woolf} 的 95% 置信限为 1.66～3.08，也与校正前 OR 的置信限很接近。最后根据公式 10－40 计算 $\chi^2_{4-1}=3.402<\chi^2_{3,0.95}=7.815$，因此 $P>0.05$。表明本组资料中，性别、年龄及 SBP 之间不存在相互作用。

10.5　相关与回归分析

相关(correlation)分析与关联分析一样，也是统计生物信息学的常用方法之一。基因组学研究中，两者都用于分析某种基因型或遗传标志与表型(疾病或性状)的关系，但两者所适用的资料类型有所差别，回归分析的数据来源于对同一个体不同变量的观察，实验设计可以是队列研究、病例对照分析或横断面分析(cross－sectional design)。相关分析常与回归分析(regression)一起使用。相关与回归分析都是研究变量之间的相互关系。相关分析主要用于描述变量之间相关性的强度；而回归分析主要是通过建立一个变量间相互关系模型，在给定某个或某些变量值的条件下，用于估计或预测其他变量值，即分析变量间的相互依存关系。

10.5.1　线性相关与回归[2～5]

描述两个变量间直线关系的密切程度和相关方向的统计指标是相关系数，通常用 r 表示，其取值范围是－1 和＋1 之间。当 $r=-1$ 时，表示完全负相关；当 $r=1$ 时，表示完全正相关；当 $r=0$ 时，称为零相关，表示不存在线性相关(不排除非线性相关关系)。相关系数的绝对值愈趋近于 1，表示直线相关程度愈密切；其绝对值愈趋近于 0 时，说明直线相关关系愈不密切。相关系数的计算是根据离均差平方和以及离均差积和等统计指标进行的。由于涉及到较大样本时，按照计算公式非常烦琐，主要利用常用统计软件(如 SPSS 等)直接给出，同时可得到该相关系数显著性检验结果(t 检验)。如果已知总体的相关系数 β，需要比较某样本的相关系数 r 是否来自同一总体，或者需要比较两相关系数是否来自同一总体，则需要先转换为 Z 值(标准正态分布)，再进行 u 检验。

线性回归是处理两变量间的线性依存关系的统计方法。通常以 x 表示自变量(independent variable)，其数值是通过实验已确定的，一般认为没有误差或误差很小。y 是依存变

量或称因变量(dependent variable),因 x 的变化而改变,属于随机变量,服从于正态分布。回归分析的目标是建立 y 与 x 的线性回归方程,使得当 x 为某一定值时,y 在某一范围内变动。即满足下列回归方程:

$$y = a + bx \qquad \text{(公式 10-41)}$$

回归方程中,y 表示由 x 所推算的估计值;a 为作图中 y 轴截距;b 为斜率,又称回归系数,表示 x 每增加或减少一个单位,y 将随之增加或减少 b 个单位。b 值可为正数,也可为负数。所得到的回归直线各点应该是因变量实际值与估计值之差的平方和的最小值,符合最小二乘法(least square method)原理。回归方程的得到主要是计算 a 和 b 值,其计算过程烦琐。实际应用中,通常也是通过软件直接得到,并进行回归方程和回归系数的显著性检验。回归方程的显著性检验是用方差分析($F_{1,n-2}$ = 回归均方/误差均方),回归系数的显著性检验是用 t 检验(自由度 $= n-2$)。

相关与回归都是用来研究变量之间的相互关系,两者既有区别,也有联系。首先,相关分析中的两个变量都是随机变量,且均服从于正态分布,主要反映两变量间的相互关系。而回归分析中,仅因变量为随机变量,自变量是固定的非随机变量,用来反映两变量间的依存关系。其次,回归分析除反映两变量间的关联性外,还可通过自变量预测因变量的数值。两者的相似之处是,相关系数与回归系数的正负号相同,即其关联性的方向相同,而且两者可以通过公式互相换算。r^2 就是指回归平方和在总平方和中所占的比值,取值范围为 0~1,其数值愈趋近于 1,表明回归效果愈好,或关联强度愈大。当 $r^2 = 0.25$ 时,表示一变量的变异中 25% 是由另一变量所引起,另外 75% 的变异是由其他因素引起。

利用 SPSS 软件进行回归分析的基本步骤是:①将数据作散点图,观察是否具有直线趋势,可同时对误差大小作图(* ZRESID vs * ZPRED),观察其分布;②建立统计假设(单侧或双侧检验);③作回归直线;④根据方差分析结果,评价回归方程的意义;⑤分析假设检验结果,作统计推断和结论。

10.5.2　协方差分析[2~5]

协方差分析(analysis of covariance, ANCOVA)是将直线回归分析与方差分析(参见 10.3.2 节)联合起来的一种统计方法,常用于消除混杂因素的影响,以提高分析结果的可靠性。其检验方法是将与 y 呈直线关系的 x 值化成几等分后,再来检验 y 修正均数间的显著性,即把各种处理所得的因变量 y 调整为一个公共的自变量 x 所应该具有的估计量。ANCOVA 分析的宗旨是利用混杂因素如性别或年龄,将自变量分为更加均匀的不同组别,分别建立回归方程,确定其显著性,同时比较各回归系数是否来自同一总体,即是否呈平行线。协方差分析可用于两组或多组资料的比较。协方差分析的假设是回归分析和方差分析假设的合并,包括:

各组样本是从具有相同方差的正态分布总体中抽样得到;

各组样本的回归方差均具有显著性意义,而且斜率相同(呈平行线)。

ANCOVA 分析的结果,如接受 H_0 假设,表示该混杂因素对因变量无影响,可以合并分

组，建立总的回归方程。

10.5.3 多元线性回归[2-5]

多元线性回归（multiple linear regression）是用线性方程来描述和分析一个因变量与多个自变量之间的数量依存关系，其表达式为：

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \qquad \text{（公式 10-42）}$$

上式中 b_0 为截距，$b_1, b_2, \cdots, b_k$ 称为偏回归系数（partial regression coefficient）。与简单线性回归一样，其计算多通过 SPSS 或 SAS 软件直接给出，同时，回归方程的显著性检验用方差分析（$F_{k,n-k-1}$ = 回归均方/误差均方），回归系数的显著性检验用 t 检验（自由度 $= n-k-1$）。与多元回归分析密切相关的是多元相关系数的计算，也称复相关系数，用 R 表示。以 R^2 表示复相关指数或确定系数，来反映自变量对因变量变异性的贡献率。在利用 SPSS 软件进行多元回归分析时，常利用散点图观察分布趋势，并对误差大小作图（*ZRESID vs *ZPRED），观察其分布。各自变量的方差，即 F - Statistic，是多元回归分析的重要参数，实际是偏回归系数 t 检验的结果换算而来：

$$F_i \approx t^2 = (b_i/SE_i)^2 \qquad \text{（公式 10-43）}$$

上式中，F_i 表示各自变量的 F - Statistic，b_i 为偏回归系数，SE_i 为该偏回归系数的标准误。

一般而言，当自变量越多，回归平方和越大，剩余平方和越小，y 估计值的误差越小，因而该回归方程的预报越精确。多元线性回归所包含的自变量较多时，有些自变量对因变量 y 的影响不显著时，需要加以剔除。一个优化的回归方程，既要考虑不能遗漏对 y 有显著影响的任何自变量，又要尽量将那些对 y 无显著性影响的变量完全排除掉，使最终的回归方程中只含对 y 影响显著的变量。优化多元线性回归方程的方法分前进法（forward selection）、后退法（backward selection）和逐步法（stepwise selection）3 种。以逐步法较常用，又称逐步回归分析。SPSS 软件中可同时提供上述 3 种选项，读者可结合实例，比较 3 种方法所得到的回归模型的差别。无论选择那种优选方法，最后模型中的结论应该相同或相似。

前进法中，首先从 b_0 项开始，界定需要引入的自变量个数。分别引入每个自变量，得到各自的回归方程，进行方差分析，确定其显著性，并定义 F - Enter 值（通常为 $P = 0.05$ 时的 F 值）。对于 F - Statistic 大于 F - Enter 的自变量，比较 F - Statistic 值大小，选择其中数值最大者，引入回归方程并固定下来，作为 b_1x_1 项。然后从 $y = b_0 + b_1x_1$ 开始，按同样的方法选定 b_2x_2，b_3x 等，直至 b_kx_k，即没有自变量满足可进入回归方程的 F - Enter 界值为止。

后退法中，先将所有 k 个变量全部引入，得到总的回归方程，界定 F - Stay 值（通常为 $P = 0.051$ 时的 F 值，与上述 F - Enter 值一样，SPSS 软件中可使用其设定的缺损值），计算各自变量的 F - Statistic 值，在 F - Statistic 小于 F - Stay 界值的变量中，选择其中 F - Statistic 值最小者，加以剔除。这样，回归方程中仅含 $k-1$ 个自变量。按同样的方法，每次剔除小于

F – Stay 值的变量中 F – Statistic 值最小者，直至不能剔除为止。

逐步法中，可视做上述两种方法相结合应用的方法。先按前进法，逐个引入自变量，每引入一个自变量后，按后退法计算各自变量的 F – Statistic 值，使之满足 F – Statistic $> F$ – Stay，否则予以剔除其中 F – Statistic 值最小者。最终回归方程中，所有自变量必须同时满足 F – Statistic 大于 F – Enter 和 F – Stay 的界定值。

多元回归分析中，可将各混杂因素和相互作用作为单独的自变量引入回归方程，然后进行显著性检验（可用方差分析比较增加变量后的回归方程与未增加变量的回归方程的 F 值），或利用逐步回归分析方法决定是否剔除该自变量。

【多元线性回归分析应用举例】

表 10 – 13 是一项有关收缩压（SBP）与相对体重（RW）的相关性调查结果，需要建立一个多元回归模型反应 SBP（因变量）与 RW 和性别的关系。

表 10 – 13　相对体重（RW%）与收缩压（SBP，mmHg）的关联性调查结果

RW%	SBP≥165	SBP < 165	合计
≥130	63	59	122
< 130	248	1 025	1273
合计	311	1 084	1 395

利用 SPSS 或 SAS 软件得到的回归方程为：

$$y = 0.159\ 58 + 0.070\ 75x_1 + 0.299\ 98x_2$$

其中，当 SBP≥165 时，$y = 1$，否则 $y = 0$；

若为女性，$x_1 = 1$，若为男性，$x_1 = 0$（x_1 均数 $= 0.524\ 731\ 2$）；

当 RW≥130，$x_2 = 1$，否则 $x_2 = 0$（x_2 均数 $= 0.087\ 455\ 2$）。

（1）根据上述多元回归方程和各自变量的均数，可得到 SBP≥165 的流行率为：

$$y = 0.159\ 58 + 0.070\ 75 \times 0.524\ 731\ 2 + 0.299\ 98 \times 0.087\ 455\ 2 = 0.2229\ 3$$

这与根据表 10 – 13 直接计算的 SBP≥165 的流行率完全相同（311/139 5 = 0.222 93）。

（2）男性和女性 SBP≥165 的流行率分别为：

$$y_{男性} = 0.159\ 58 + 0.299\ 98 \times 0.087\ 455\ 2 = 0.185\ 8$$

$$y_{女性} = 0.159\ 58 + 0.070\ 75 + 0.299\ 98 \times 0.087\ 455\ 2 = 0.256\ 56$$

即女性 SBP≥165 的流行率高于男性，其相对风险率 RR 为：0.256 56/0.185 8 = 1.38。

（3）不同 RW 水平 SBP≥165 的流行率分别为（性别已校正）：

$$y_{RW \geq 130} = 0.159\ 58 + 0.070\ 75 \times 0.524\ 7312 + 0.299\ 98 = 0.496\ 68$$

$$y_{RW < 130} = 0.159\ 58 + 0.070\ 75 \times 0.524\ 7312 = 0.196\ 7$$

校正性别后，RW≥130 与 RW < 130 相比较，高 SBP 的相对风险率 RR 为：0.496 68/0.196 7 = 2.525，这与校正前根据表 10 – 13 中资料，按式 10 – 19 的直接计算结果：（63 × 1 273）/（248 × 122）= 2.65 有所区别。由此可见，多元回归方程可直接根据自变量预测因变量的变化，并对混杂因素进行校正。

10.5.4 Logistic 回归[2,4]

10.5.4.1 简单 Logistic 回归

Logistic 回归(Logistic regression)是回归分析中的一种特殊类型,遵循回归分析的基本规律,适用于离散变量的分析,主要用来预测发病率和几率比值的计算。其简单回归方程可表示为:

$$\ln Odds = \ln[P/(1-P)] = \alpha + \beta X \qquad (公式\ 10-44)$$

上式中,P 表示发病率,α 表示人群或样本中的发病风险,β 表示 X 每改变一个单位时的 lnOR 值,即 OR 等于 β 的反对数值 e^{β}。例如表 10-14 为血清胆固醇水平与冠心病相关性的调查结果。

表 10-14 血清胆固醇(Chol)与冠心病(CHD)的关联性调查结果

Chol (mmol/L)	CHD 患者		合计
	是	否	
≥6.72	91	295	386
<6.72	177	800	977
合计	268	1095	1363

根据上表资料,在 Logistic 分析中,对 X 的定义如下:

当 Chol≥6.72 mmol/L,$X=1$,否则 $X=0$。

通过 SAS 软件得到的 Logistic 回归方程为:$\ln Odds = -1.507 + 0.329X$。以下比较几种方法得到的 OR 值。

(1) 根据对 X 的定义和 Logistic 方程计算 OR:

当 $X=1$ 时,$\ln Odds = -1.507 + 0.329 = -1.178$,$OR_{\geq 6.72} = e^{-1.178} = 0.308$;

当 $X=0$ 时,$\ln Odds = -1.507 + 0 = -1.507$,$OR_{<6.72} = e^{-1.507} = 0.222$;

$OR = OR_{\geq 6.72} / OR_{<6.72} = 0.308/0.222 = 1.39$。

(2)根据 Logistic 方程中的 β 值计算 OR:

$OR = e^{\beta} = e^{0.329} = 1.39$。

(3)根据公式 10-22 计算 OR:

$OR = (91 \times 800)/(295 \times 177) = 1.39$。

同样,利用 Logistic 方程可以计算出 CHD 的发病率:

当 $X=1$ 时,$P/(1-P) = 0.308$,因此 $P_{\geq 6.72} = 0.235$;

当 $X=0$ 时,$P/(1-P) = 0.222$,因此 $P_{<6.72} = 0.181$。

利用表 10-14 资料,可得到相同结果,读者可自行验证。

10.5.4.2 多元 Logistic 回归

对于多元 Logistic 回归分析,其回归方程可表示为:

$$\ln Odds = \ln[P/(1-P)] = \alpha + \Sigma\beta_i X_i \qquad (公式\ 10-45)$$

$$P = 1/[1 + e^{-(\alpha + \Sigma\beta_i X_i)}] \qquad (公式\ 10-46)$$

同样可利用公式10-45和10-46分别计算出各OR值、总OR值和发病率P。上述公式中β_i表示在其他变量固定条件下，当X_i每改变1个单位时，对ln*Odds*的改变量；如果改变10个单位，对OR的变化是$e^{10\beta}$。如果两个变量同时改变，而其余变量相同，这时对OR的变化等于$e^{\beta_1+\beta_2}$。对于OR的置信区间的计算，可根据公式10-47。

$$95\%\,CL\,(\ln OR) = \ln OR \pm 1.96SE(\ln OR) = \beta \pm 1.96SE(\beta) \qquad (公式\ 10-47)$$

Logistic回归分析中，对于两分变量(dichotomous variable)，通常定义1="是"，0="否"。对于多阶变量或连续变量，须转换为亚变量(dummy variable)。如对于收缩压变量(SBP)，可拆分为165、140等不同界值，因此得到不同的SBP变量：

当SBP≥165 mmHg，则SBP1=1，否则SBP1=0；

当SBP≥140但<165mmHg，则SBP2=1，否则SBP2=0；

当SBP<140 mmHg，则SBP3=1，否则SBP3=0。

对于混杂因素或相互作用，同样可以利用多元Logistic回归进行剔除或引入。判断某变量是否引入的方法是根据-2Log似然性计算结果(SAS软件自动给出)，两个不同方程中-2Log似然性的差值即为χ^2值，自由度等于不同回归方程中自变量数目之差。这种校正方法较10.4.5节中所述的几率比值校正方法要简便，尤其是考虑多个混杂因素或相互作用的情况。

【多元Logistic回归分析应用举例】

例如，通过18年连续观察1 363例研究对象，以分析冠心病(CHD)发病率与收缩压(SBP)、血清胆固醇(Chol)、年龄、性别的关系，利用SAS软件得到以下Logistic回归方程：

$$P(CHD) = y = 1\{1 + \exp[-(-8.339\,8 + 0.051\,6\mathrm{Age} + 1.013\,2\mathrm{Sex} + 0.016\,8\mathrm{SBP} + 0.004\,76\mathrm{Chol})]\}$$

其中，当诊断为冠心病时，$y=1$，否则$y=0$；

Age表示实际年龄；

若为男性，Sex=1，女性Sex=0；

SBP以实测mmHg值表示；

Chol以实测血清胆固醇浓度(mg/dL)表示。

(1) 利用标准化β_i值(以β_i^*表示，$\beta_i^* = \beta_i \times SD$)，分析其中对CHD发病率影响最大的自变量。SAS分析结果见表10-15。

表 10－15 β_i值标准化有关统计量表

变量(X_i)	β_i	$SE(\beta_i)$	$SD(X_i)$	β_i *	排序
截距	－8.339 8	0.950 7	－	－	－
Age	0.051 6	0.015 2	4.784 2	0.246 9	3
Sex	1.013 2	0.152 9	0.499 4	0.506 0	1
SBP	0.016 8	0.002 46	27.808 5	0.467 2	2
Chol	0.004 76	0.001 57	46.352 8	0.220 6	4

表 10－15 中 β_i * 数值代表当变量每改变一个标准差时，CHD 的 ln*Odds* 值改变大小。结果表明，该组变量中，Sex 变量对发病风险影响最大，而 Chol 影响最小。

（2）计算一名 60 岁女性，SBP＝185 mmHg，Chol＝277 mg/dL，其 18 年期 CHD 的发生率为多少？将有关数值代入上述回归方程中：

$$P(\mathrm{CHD}) = 1/\{[1+\exp-(-8.339\ 8+0.051\ 6\times 60+1.013\ 2\times 0+0.016\ 8\times 185+0.004\ 76\times 277)]\}$$

$$=1/[1+\exp(0.817\ 3)]=0.306$$

同样参数值的男性，其预期 CHD 发病风险率为 0.549（读者可自行计算），较女性的发病风险明显增加。

（3）如果 Sex，Age，Chol 3 个变量固定，SBP 差值为 25 mmHg，其 OR 值为多少？

$$OR = e^{0.0168\times 25} = e^{0.42} = 1.52$$

（4）为了分析 Sex 和 Age 是否为此回归模型的混杂因素，通过 SAS 分析得到表 10－16 数据（各自变量的定义同上）。

通过比较模型 2 和模型 4 的－2Log 似然性数值差：1 301.761－1 243.516＝58.245，自由度＝4－2＝2，$\chi^2_{2,0.95}$值＝5.99，因此 $P<0.05$。即认为 Sex 和 Age 是混杂因素。

表 10－16 不同回归模型的回归系数及－2*Log* 似然性数值

自变量(X_i)	模型 1	模型 2	模型 3	模型 4
截距	－3.769 1	－4.338 2	－6.801 7	－8.339 8
SBP	0.015 6	0.015 1	0.013 9	0.016 8
Chol		0.002 71	0.002 49	0.004 76
Age			0.051 2	0.051 6
Sex				1.013 2
－2Log 似然性	1 305.048	1 301.761	1 289.956	1 243.516

最后，我们将几种常用的多元分析方法的比较总结如表 10－17。

表 10-17 几种常见的多元分析方法的比较

分析方法	变量类型		分析目的
	因变量	自变量	
方差分析(ANOVA)	连续变量	分类变量	描述组间变异是否具有显著性
协方差分析(ANCOVA)	连续变量	分类变量(主变量) 连续变量(混杂因素)	描述考虑混杂因素情况下,组间变异的差异显著性
多元线性回归	连续变量	传统上为连续变量,实际应用中也可为离散变量或分类变量	预测和描述自变量与因变量间的关联强度与方向
多元 Logistic 回归	两分变量	连续或分类变量	预测和描述一个或多个自变量与某事件(疾病)发生概率的关系

多元回归分析中,通常采用 Hosmer-Lemeshow 的策略,首先进行单变量分析,找出与因变量相关性有显著意义的变量(设定 $P<0.25$ 为选择标准)。然后将所有单变量分析中选定的变量代入多元线性回归或多元 Logistic 回归模型中,用后退法建立初步多元回归模型。引入所有可能的相互作用变量,再用后退法筛选变量,得到最后的多元回归模型。从统计学和生物学等角度评价回归模型中的各自变量或相互作用变量的意义。相互作用变量通常在以下情形时需要考虑:治疗方法与疾病严重度、年龄与性别或种族之间、年龄与疾病类型、测定时受试者的状况与测定方法、日期与治疗方法、不同症状之间等。一个合适的多元回归模型尚有待于不同试验的验证。

10.6 常用统计软件介绍

10.6.1 SPSS 软件[6]

SPSS 软件是由美国 SPSS 公司 (http://www.spss.com)设计的统计专用软件。SPSS 公司在中国也设立了分公司,目前市面上已有中文版 SPSS 软件和操作说明书。这里仅简单介绍 SPSS 的基本使用方法,读者需要在购买和安装软件后,参考操作手册或软件的帮助菜单,结合具体资料,反复练习,才能熟练掌握。

SPSS 软件可在 Windows 环境下操作,也可在 Unix 或其他操作环境中运行。这里仅以 Windows 支持下的 SPSS 10.0 英文版软件为例。启动 SPSS 软件一般在开始菜单的运行程序中点击进入,直接进到数据编辑器(Data Editor)视窗。即可按读者要求输入数据,存档建立相应的数据文件,也可打开已有的数据文件。数据文件的扩展名为 sav。数据编辑器的最上面一行为变量名,可从左下方的 Data View 窗口切换到 Variable View 窗口,然后定义各变量的相关参数。回到 Data View 窗口在各栏中输入相应数据。如须对原始数据进行转换,可进入"转换(Transform)"菜单作适当转换,或产生新的变量。另外 SPSS 也可读取 Lotus、

Excel 等数据处理器存储的文档，这也需要利用合适的命令或利用“Import Wizard”转换为 SPSS 可读文件。

数据输入后，可在“分析（Analyze）”菜单中作基本统计量、相关、回归、聚类、非参数统计、生存分析等。逐级进入各统计方法的子菜单进行所需要的统计分析。统计分析前，往往跳出一菜单，需要选择所分析的变量，设置相应的统计参数，并在各高级选项设置作图方法、输出参数等。任何情况下，若对某种用法有疑问，可随时点击帮助菜单，进行求助。程序运行结束后，会出现输出结果的显示屏，可按读者需要打印所需页码或对输出结果存档。输出显示中，左边框为目录，方便用户查阅所需结果。

SPSS 的另一常用菜单是“图形（Graph）”菜单，可进行各种不同的图形处理。SPSS 使用结束后，点击 Exit 退出该程序。

相对 SAS 软件而言，SPSS 软件在学习使用上要简便和直观，称之为 user - friendly。如何用好 SPSS 软件，除熟悉其操作方法外，关键还是统计学的基础。SPSS 软件在图形处理上也较 SAS 软件方便。究竟使用 SPSS 好，还是 SAS 好，需要根据用户的喜好、统计类型、软件购买能力等进行选用。

10.6.2 SAS 软件[7]

SAS 软件是美国 SAS 软件公司设计出的大型统计软件，在生物统计、公共卫生、金融业、保险业、市场营销等领域应用都非常广泛。这是一个非常齐全的软件，要完全熟悉 SAS 的所有操作方法，需要花时间浏览几英寸厚的操作指南。这对大多数用户而言，是个令人沮丧的事情。不妨用到哪、学到哪，特别是借助 SAS 的帮助菜单，积少成多，最终就会喜欢和离不开 SAS 了（当然不是 SARS!）。

与 SPSS 一样，SAS 既可在 Windows 支持下运行，也可在 Unix 等环境中运行。这里以 Unix 支持下的 SAS 8.0 英文版为例，对其基本操作和应用技巧作以简单介绍。

SAS 的启动是在开始菜单的运行程序中点击进入。会出现几个不同的窗口，用户按自己喜好作层叠或平铺排列。这些窗口包括程序编辑器（program editor）、日志（log）、输出（output）等。其菜单功能相对简单，易于掌握。SAS 应用中最关键的步骤当然是编写程序。这需要在熟悉资料类型的基础上，根据统计目的，利用 SAS 语言编写（须掌握基本的 SAS 语言语法和句法）。

SAS 编程中需要注意以下几个问题：

（1）SAS 程序运行中，与字母大小写和空格数无关，是否使用纯粹为了程序直观和根据个人喜好。

（2）SAS 程序因句法错误，可导致运行失败。随时检查日志中的出错信息，有助于找到错误之处。最常见的错误是分号“;”的使用（不该用或没有用），其次是误将单引号打成双引号，再次是忘了加“RUN”语句或空格分隔，还有就是在变量名之间未用空格分开。

（3）小的 SAS 数据文件的产生可用 SAS 编写，大数据更常用的是由 Excel 或 Text 等文件转换而来。

以下先以一个练习题为例，列举 SAS 程序编写、运行日志和相应输出结果，说明其基本

应用过程,括号内的中文为解释性加注。

SAS LOG of EXAM 1（SAS 日志）

```
1
dm 'output; clear; log; clear;';（清屏命令）
2  libname Exam1 '/home/grad/rpn/wli8';（须指定 libname 名称如 Exam1 及其相应路径）
NOTE: Libref EXAM1 was successfully assigned as follows:（运行状况）
      Engine:        V8
      Physical Name: /home/grad/rpn/wli8
3  DATA ONE;（编辑新的工作数据文件 WORK.ONE）
4  set Exam1.Fram30;（源数据文件 Exam1.Fram30）
5  TITLE 'FRAMINGHAM HEART STUDY - 30 YEAR FOLLOWUP';（加标题 1）
6  TITLE2 'SPM 506 EXAM 1 DATA SET';（加标题 2）
NOTE: There were 5209 observations read from the data set EXAM1.FRAM30.
NOTE: The data set WORK.ONE has 5209 observations and 5 variables.
NOTE: DATA statement used:
      real time          2.49 seconds
      cpu time           0.24 seconds
7  PROC CONTENTS;（查看内容）
8  RUN;（程序运行命令）
NOTE: PROCEDURE CONTENTS used:
      real time          0.86 seconds
      cpu time           0.07 seconds
9  PROC FREQ;（查看频数）
10  TABLES SEX CHD;（列出 SEX, CHD 两个变量的频数结果）
11  RUN;（程序运行命令）
NOTE: There were 5209 observations read from the data set WORK.ONE.
NOTE: PROCEDURE FREQ used:
      real time          0.91 seconds
      cpu time           0.04 seconds
12  PROC MEANS N NMISS MEAN MIN MAX;（计算平均数：列出样本数 N，缺损样本数，
平均数，最小值，最大值）
13  VAR AGX1 BMI;（变量 AGX1 和 BMI）
14  RUN;（程序运行命令）
NOTE: There were 5209 observations read from the data set WORK.ONE.
NOTE: PROCEDURE MEANS used:
      real time          0.87 seconds
      cpu time           0.05 seconds
```

```
15  DATA Two;(编辑新的工作数据文件 WORK.TWO)
16  set Exam1.Fram30;(源数据文件 Exam1.Fram30)
17   * Definition of BMI groups;(定义新的分组变量 BMI1)
18  IF BMI = . Then delete;(“.”表示数据缺失)
19  IF BMI >= 30 THEN BMI1 = 1;
20  IF BMI >= 25 and BMI < 30 THEN BMI1 = 2;
21  IF BMI < 25 Then BMI1 = 3;
NOTE: There were 5209 observations read from the data set EXAM1.FRAM30.
NOTE: The data set WORK.TWO has 5199 observations and 6 variables.
NOTE: DATA statement used:
      real time        1.55 seconds
      cpu time         0.18 seconds
22  proc format;(格式命令)
23  value bmi1fmt 1 = '>=30'  2 = 25-29'  3 = '<25';
NOTE: Format BMI1FMT has been output.
NOTE: PROCEDURE FORMAT used:
      real time        2.06 seconds
      cpu time         0.04 seconds
24  proc freq;(频数统计-列表)
25  table bmi1/nopercent norow nocol;(数据表格格式)
26  FORMAT BMI1 bmi1fmt.;(排列方式,注意 bmilfmt 后有一圆点)
27  run;(程序运行命令)
NOTE: There were 5199 observations read from the data set WORK.TWO.
NOTE: PROCEDURE FREQ used:
      real time        0.47 seconds
      cpu time         0.03 seconds

28  DATA THREE;(编辑新的工作数据文件 WORK.THREE)
29  set TWO;(源数据文件为 WORK.TWO)
30   * Definition of CHD groups;(定义新的 CHD 变量)
31  IF CHD = 1 Then CHD1 = 1;
32  IF CHD > 1 Then CHD1 = 2;
33  IF CHD = 0 Then CHD1 = 3;
34  IF CHD1 = 1 Then CHD2 = .;
35  IF CHD1 = 2 Then CHD2 = 1;
36  IF CHD1 = 3 Then CHD2 = 2;
37   * Definition of new BMI variable;(定义新的 BMI 变量)
```

```
38  IF BMI1 = 1 THEN BMI2 = 1;
39  IF BMI1 = 2 THEN delete;
40  IF BMI1 = 3 THEN BMI2 = 2;
NOTE: There were 5199 observations read from the data set WORK. TWO.
NOTE: The data set WORK. THREE has 3032 observations and 9 variables.
NOTE: DATA statement used:
      real time          2.05 seconds
      cpu time           0.23 seconds
41  proc format; (格式命令)
42  value chd1fmt 1 = 'prevalent'  2 = 'incident'  3 = 'negative';
NOTE: Format CHD1FMT has been output.
43  value chd2fmt 1 = 'yes'  2 = 'no';
NOTE: Format CHD2FMT has been output.
44  proc format;
45  value bmi2fmt 1 = '> =30'  2 = '<25';
NOTE: Format BMI2FMT has been output.
NOTE: PROCEDURE FORMAT used:
      real time          1.73 seconds
      cpu time           0.00 seconds
46  proc freq;
47  tables BMI2 * CHD2/nopercent nocol norow CMH; (几率比值运算)
48  FORMAT BMI2 bmi2fmt. CHD2 chd2fmt. ; (表格格式,注意圆点的存在)
49  run;
NOTE: There were 3032 observations read from the data set WORK. THREE.
NOTE: PROCEDURE FREQ used:
      real time          1.23 seconds
      cpu time           0.06 seconds
```

SAS OUTPUT of EXAM 1 (SAS 运算结果)

```
                    FRAMINGHAM HEART STUDY - 30 YEAR FOLLOWUP
                              SPM 506 EXAM 1 DATA SET
                          10:27 Saturday, February 22, 2003

                              The CONTENTS Procedure

Data Set Name: WORK. ONE              Observations:          5209
Member Type: DATA                     Variables:             5
```

Engine: V8 Indexes: 0
Created: 10:31 Saturday, February 22, 2003 Observation Length: 40
Last Modified: 10:31 Saturday, February 22, 2003 Deleted Observations: 0
Protection: Compressed: NO
Data Set Type: Sorted: NO
Label:

The CONTENTS Procedure
(数据文件内容)
- - - - - Alphabetic List of Variables and Attributes - - - - -

#	Variable	Type	Len	Pos	Label
1	AGX1	Num	8	0	AGE, EXAM 1
5	BMI	Num	8	32	BODY MASS INDEX(kg/m * *2), Exam 1
2	CHD	Num	8	8	FIRST EVIDENCE OF CHD
3	DTH	Num	8	16	DEATH
4	SEX	Num	8	24	SEX

The FREQ Procedure
(频数统计)

SEX

SEX	Frequency	Percent	Frequency	Cumulative Percent
1	2336	44.85	2336	44.85
2	2873	55.15	5209	100.00

FIRST EVIDENCE OF CHD

CHD	Frequency	Percent	Frequency	Cumulative Percent
0	3760	72.18	3760	72.18
1	82	1.57	3842	73.76
2	59	1.13	3901	74.89
3	56	1.08	3957	75.96
4	76	1.46	4033	77.42
5	66	1.27	4099	78.69
6	85	1.63	4184	80.32
7	81	1.56	4265	81.88

8	103	1.98	4368	83.85
9	92	1.77	4460	85.62
10	105	2.02	4565	87.64
11	139	2.67	4704	90.31
12	120	2.30	4824	92.61
13	95	1.82	4919	94.43
14	95	1.82	5014	96.26
15	90	1.73	5104	97.98
16	105	2.02	5209	100.00

The FREQ Procedure

（频数统计表）

Table of BMI2 by CHD2

BMI2	CHD2		
Frequenc y	yes	no	Total
> =30	270	470	740
<25	465	1827	2292
Total	735	2297	3032

The FREQ Procedure

Summary Statistics for BMI2 by CHD2

Cochran – Mantel – Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	79.8978	<0.0001
2	Row Mean Scores Differ	1	79.8978	<0.0001
3	General Association	1	79.8978	<0.0001 (显著性检验)

The FREQ Procedure

Summary Statistics for BMI2 by CHD2

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits

Case – Control (Odds Ratio)	Mantel – Haenszel	2.2571	1.8834	2.7050	(OR 结果 1)
	Logit	2.2571	1.8834	2.7050	(OR 结果 2)
Cohort (Coll Risk)	Mantel – Haenszel	1.7984	1.5871	2.0379	(RR 结果)
	Logit	1.7984	1.5871	2.0379	
Cohort (Col2 Risk)	Mantel – Haenszel	0.7968	0.7516	0.8447	
	Logit	0.7968	0.7516	0.8447	

(备注：OR 结果 1 是用 Mantel – Haenszel 法，详见 10.4.5.1 节；OR 结果 2 是用 Woolf 法，详见 10.4.5.2 节)

以上例子中包括最基本的统计分析方法，如基本统计量分析、RR 与 OR 分析等，下面将其他一些关联或回归分析中常用的 SAS 统计命令列出，供参考。

(1)OR 或 RR 分析中须对混杂因素进行校正：

proc freq; tables sex * age * sbp * CHD/CMH alpha = 0.05;

此命令用于分析 sbp 与 CHD 关联分析时，计算校正 age 和 sex 后的 OR 或 RR 值。

(2)线性相关与回归：

proc glm; model y = x1 x2; (协方差分析)

proc reg simple; model y = age x1 x2 age_x1 age_x2; plot r. * p.; [多元线性回归中考虑混杂因素和相互作用变量 age_x1 和 age_x2，同时对剩余值(纵轴)与预期值(横轴)作图]

(3)Logistic 回归：

proc logistic; model y = age sex sbp chol; (当 y = 0, yes; y = 1, no)

proc logistic decending; model y = age sex sbp chol; (当 y = 1, yes; y = 0, no)

proc logistic simple; model y = age sex sbp chol/LACKFIT; (列出回归模型的观察值与预期值的比较，以评价回归模型的好坏)

(李 巍)

参 考 文 献

1 Guttman I, Wilks SS, Hunter JS. Introductory Engineering Statistics. 3rd ed, New York: John Wiley & Sons, 1982.5 ~ 150

2 Daniel WW. Biostatistics: A Foudation for Analysis in the Health Sciences. New York: John Wiley & Sons, 1999. 83 ~ 570

3 江三多，吕宝忠. 医学遗传数理统计方法. 北京：科学出版社，1998. 213 ~ 233

4 Kahn HA, Sempos CT. Statistical Methods in Epidemiology. Oxford: Oxford University Press, 1989.45 ~ 165

5　刘筱娴. 医学统计学. 北京：科技出版社，2000. 224～249
6　SPSS Base 10.0 Brief Guide. Chicago：SPSS，2000. 1～58
7　Delwinche LD，Slaughter SJ. The Little SAS Book. 2nd ed. Champl Hill:SAS，1996. 25～89